Redbooks
ibm.com/redbooks

Intelligent Storage Management with Artificial Intelligence for IT Operations (AIOps) using IBM Storage Insights

Archana Chinnaiah
Binayak Dutta
David Green
Deepak Salian
Keigo Matsubara
Ramakrishna Vadla
Sanjay Lokhande
Sumesh Pahuja
Sushil Sharma
Utkarsh Singh
Vasfi Gucer
Vikash Bhati

Storage

Artificial Intelligence

IBM

Redpaper

IBM Redbooks

Intelligent Storage Management with AIOps using IBM Storage Insights

May 2025

REDP-5755-00

Note: Before using this information and the product it supports, read the information in “Notices” on page v.

First Edition (May 2025)

This edition applies to the IBM Storage Insights (a software as a service (SaaS) product) Q1 2025 release.

This document was created or updated on May 13, 2025.

Contents

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

Preface

This IBM® Redpaper delves into the practical application of AI and automation within IBM Storage Insights, empowering you to optimize your storage environment through data-driven insights. From accurate capacity planning to detecting performance deviations and leveraging AI-based workload advisories for storage partitions, this guide equips you with the tools to master proactive storage management. We will explore advanced capabilities such as system health advisories for IBM FlashSystem® grid, ransomware threat detection, the power of the IBM Storage Insights Observability Chatbot, and the innovative potential of artificial intelligence for storage (AI4S), all within the context of enhancing efficiency, security, and performance.

The target audience of this Redpaper is storage administrators and storage technical specialists wanting to learn about Artificial Intelligence for IT Operations (AIOps) in the storage context.

Authors

This paper was produced by a team of specialists from around the world.

Archana Chinnaiah is a Software Engineer at IBM Systems Development Lab (ISDL) in Bangalore who contributes to IBM Storage Insights. With 9 years of experience at IBM, she specializes in back-end development and design for IBM Storage Insights. She holds a master's degree in Software Engineering from PSG College of Technology, Coimbatore.

Binayak Dutta is the AI Architect and lead Data Scientist at IBM India Systems Development Labs (ISDL), where he leads the development and implementation of the AI feature roadmap for IBM Storage Insights. His mission is to leverage AI ethically to differentiate products and improve business outcomes.

David Green works with the IBM SAN Central team to troubleshoot performance and other problems on storage networks. He authored or contributed to several IBM Redbooks® publications. He is a regular speaker at IBM Technical University.

Deepak Salian is a SME with IBM Technology Lifecycle Services who specializes in IBM FlashSystem and IBM Storwize®. He holds a bachelor's degree in Electronics Engineering from the University of Mumbai India. He has over 9 years of experience in IT. Before stepping into his SME role, he served as a Resident Engineer for Reliance Jio, one of India's largest telecommunications companies. In this role, he provided onsite technical support and expert guidance to help ensure the seamless operation of critical storage systems, including IBM FlashSystem, SVC, IBM DS8000®, and IBM Tape Systems.

Keigo Matsubara is a technical sales specialist working for various storage products, including IBM Storage FlashSystem and IBM Storage Insights. He has 33 years of industry experience. He focuses on IBM Storage products such as IBM Storage FlashSystem and IBM Storage Insights. He joined IBM Japan in 1992 as a programmer and later moved to his current role.

Ramakrishna Vadla is a Senior Technical Staff Member (STSM) and Lead Architect for IBM Storage Insights and IBM Spectrum® Control. He is responsible for developing and designing the IBM Storage Insights and IBM Spectrum Control products, which monitors storage systems, fabric and servers. With over 20 years of experience, he has worked on large-scale distributed systems across various technologies, including storage management, AIOps, microservices architecture, cloud-native services, and middleware systems. He has spoken at multiple technical forums, including the SNIA Storage Developer Conference and IBM global conferences, and has contributed to the open-source community. He holds a Master of Technology degree in Computer Science from the International Institute of Information Technology, Hyderabad, India.

Sanjay Lokhande is a Development Lead at ISDL, Pune. He has 19 years of software development experience, and worked in supply chain management, cloud, and storage domain. He is the architect for IBM Spectrum® Control. He worked on delivering various features like Ceph integration, and AI suggestion in IBM Storage Insights. He holds Bachelor of Engineering degree in Computer Science and Engineering from RIT, Shivaji University, India.

Sumesh Pahuja works with the IBM Storage Insights development team as a Senior Development Manager. He has more than 24 years of experience in IT and specifically the storage domain. His overall experience spans across Development and Expert Labs. He has extensive experience with the IBM Storage portfolio, which includes IBM Storage DS8000, IBM Storage FlashSystem, and IBM Storage software.

Sushil Sharma is a Senior Software Engineer working as the Back-end Architect for IBM Storage Insights at ISDL, Pune. He has 18 years of industry experience and worked on IBM FlashSystem and SVC since 2015. He worked and delivered features, including AI-Based Capacity Planning, IBM FlashSystem grids and Migration from IBM Storage Insights, IBM Spectrum Virtualize for Azure, IBM Spectrum Virtualize for AWS, 3-Site Replication (Metro Mirror, IBM HyperSwap®, and GUI), and iSCSI and iSER support on SVC. He contributed to many customer engagement programs. He also created various customer collaterals, such as IBM Blueprints and technical white papers. He holds a master's degree in Computer Application from Mumbai University, India.

Utkarsh Singh is a Machine Learning Engineer at ISDL (ISDL), Pune, who primarily contributes to AIOps initiatives. He specializes in developing AI-driven solutions for predictive analytics, anomaly detection, and intelligent automation to enhance IT operations. His work spans workload forecasting, automated capacity prediction, and ransomware threat detection, where he optimized machine learning pipelines and real-time inference systems. He also collaborated extensively with IBM Research® Labs on applied AI solutions and actively contributed to the open-source community, notably developing a natural language chatbot for IBM Storage Insights. In addition to implementation, Utkarsh is research-driven, with multiple published papers, aligning with IBM's vision of innovation in AI and enterprise automation. He holds a Bachelor of Technology degree in Electronics and Communication Engineering from MIT World Peace University, Pune, India.

Vasfi Gucer leads projects for the IBM Redbooks team, using his 25+ years of experience in systems management, networking, and software. A prolific writer and global IBM instructor, his focus shifted to storage and cloud computing in the past 8 years. Vasfi holds multiple certifications, including IBM Certified Senior IT Specialist, PMP, ITIL V2 Manager, and ITIL V3 Expert.

Vikash Bhati is a Senior Software Engineer working as the Full Stack developer at ISDL, Pune. He has 6 years of industry experience, and has been working on IBM Storage Insights since 2022. He has worked and delivered features, including Capacity Planning, IBM Storage Insights Provisioning System, IBM FlashSystem grids, and Workload Mobility for IBM Storage Insights. He contributed to developing a ransomware threat detection AI model with the IBM Zurich Research Lab. He holds a bachelor's degree in Electrical Engineering from Gujarat Technological University, India.

Thanks to the following people for their contributions to this project:

Piyush Gupta, Randhir Singh, J Lakshmi Yoganand
IBM India

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an IBM Redbooks residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our papers to be as helpful as possible. Send us your comments about this paper or other IBM Redbooks publications in one of the following ways:

- Use the online **Contact us** review Redbooks form found at:

  ibm.com/redbooks

- Send your comments in an email to:

  redbooks@us.ibm.com

- Mail your comments to:

  IBM Corporation, IBM Redbooks
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- Find us on LinkedIn:

  https://www.linkedin.com/groups/2130806

- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/subscribe

- Stay current on recent Redbooks publications with RSS Feeds:

  https://www.redbooks.ibm.com/rss.html

1

Capacity planning

This chapter provides a high-level understanding of the following topics:

1.1 Need for capacity planning

In the modern world, data has grown. To store this data, organizations need to constantly upgrade the capacities of their storage solutions.

Organizations either underestimate or overestimate their storage needs. Last-minute scrambles might happen if the need for storage capacity is underestimated. However, wastage of resources and increases in cost might happen if storage capacity is overestimated.

Planning for your future capacity needs is an important part of ensuring that your operations keep running smoothly as your organization grows. With the rapid proliferation of applications and devices into every aspect of business, the demand for data storage has never been higher.

To help you stay ahead of demand, avoid critical capacity shortages, and make more informed purchasing decisions, you can forecast your future capacity needs based on the analysis of your historical capacity usage. By using past metrics and advanced formulas, you can project, chart, and report on future capacity usage for block storage systems, file storage systems, pools (by tier), and Storage as a Service (STaaS) environments, such as the following ones:

- Block storage systems:
  - IBM Storage FlashSystem.
  - IBM System Storage DS8000 series.
  - IBM SAN Volume Controller (SVC).
  - IBM Storage Virtualize.
  - Third-party block devices:
    - Dell EMC. For more information, see Dell EMC.
    - Hitachi. For more information, see Hitachi.
    - NetApp. For more information, see NetApp.
    - Pure Storage. For more information, see Pure Storage.
- File storage systems:
  - IBM Storage Scale.
  - IBM Storage Scale System.
- Pools (by tier).
  - A pool is a logical grouping of storage resources to monitor and analyze usage.
  - Users can create custom pools based on storage tiers, such as Performance, Capacity, or Archive. Some IBM Storage Systems (like IBM Storage Scale and IBM Storage FlashSystem) support automated data movement between tiers.

    For more information, see Tiers.
- Storage as a Service (STaaS) environments. For more information, see Storage as a Service dashboard.

1.2 Capacity forecasting by using IBM Storage Insights AIOps capabilities

IBM Storage Insights can forecast capacity usage trends and generate a graphical representation with the help of an advanced artificial intelligence (AI) model.

IBM Storage Insights uses multiple AI models to predict future capacity requirement trends and presents the best possible result to the user.

The AI models are trained on your storage system capacity historical usage trend, which means this AI model can predict future capacity usage with maximum accuracy because this AI model is trained only for your use case.

To predict the capacity usage of your storage system, log in to your IBM Storage Insights instance, hover your cursor over the menu in the left pane of the dashboard, and select **Planning** → **Capacity Planning**. The Capacity planning dashboard opens and lists all the block storage systems by default that were added to the tenant, as shown in Figure 1-1.

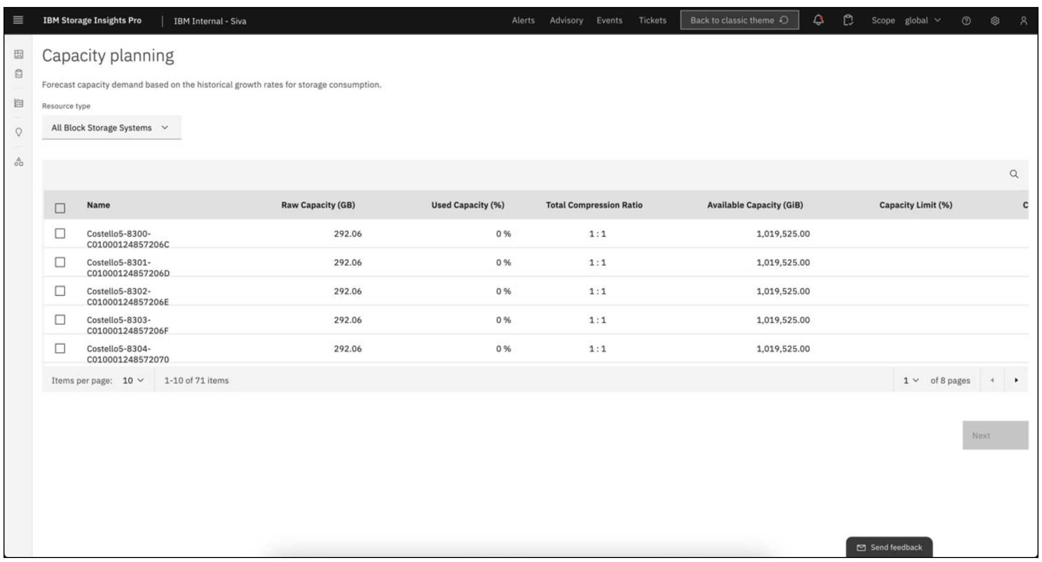


Figure 1-1 Capacity planning dashboard

By default, block storage systems are listed, but you can select other storage systems by clicking the **Resource type** drop-down menu.

In this use case, block storage systems are considered for capacity planning. Select your storage system by selecting its checkbox and click **Next** to run the prediction for future capacity usage.

Once a prediction is complete, a detailed view of future capacity usage trends is automatically presented, including a chart with historical data (up to one year prior) and a default forecast for the next year, as depicted in Figure 1-2.

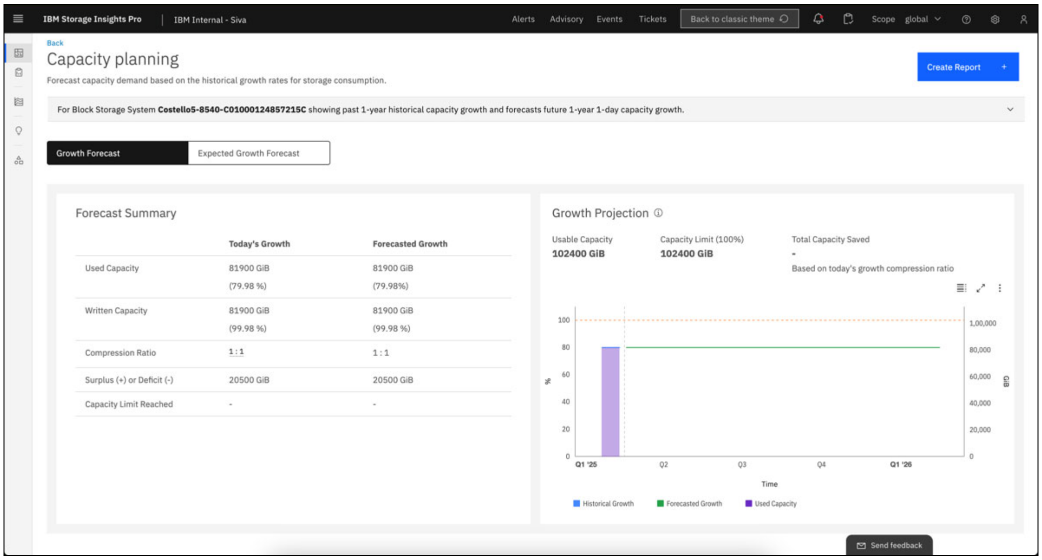


Figure 1-2 Capacity Forecast view

In Figure 1-2, under the Forecasted Growth column, different details are shown, such as Forecasted Used Capacity, Forecasted Written Capacity, Compression Ratio, Surplus (+) or Deficit (-), whether the storage system is about to reach the preset Capacity Limit, and Capacity Limit Reached Date:

- Forecasted Used Capacity

  Forecasted capacity usage is based on the growth that you expect during the selected time range. Use this value to help gauge your future capacity usage and potential capacity needs.

- Forecasted Written Capacity

  The forecasted amount of capacity during the selected time range is written to the volumes in a pool before inline disk compression is applied.

- Surplus (+) or Deficit (-)

  The amount of the capacity that is excess or deficit compared to Forecasted Used Capacity. The value can be positive or negative.

  – If the value is positive, it indicates a surplus, meaning that there is available capacity beyond the forecasted usage.

  – If the value is negative, it indicates a deficit, meaning that the forecasted usage exceeds the capacity limit, potentially leading to capacity shortages.

- Capacity Limit

  The future capacity limit of the storage system or pool (in GiB) is based on the original limit that was set. For example, if the capacity limit was set to 80% and the projected used capacity is 100 GiB, the value 80.00 GiB is displayed. If the capacity limit value is not set, then the forecast capacity limit is 100%.

- Capacity Limit Reached

  The date when the specified capacity limit for the storage system will be reached is based on historical analysis.

- Under Growth Projection, you can see the Forecasted Capacity trend:
  - The blue line of the chart shows the historical growth trend.
  - The green line shows the forecasted growth trend. This trend line is generated by AI models.
  - The purple line shows the used capacity trend to date.

  IBM Storage Insights Pro allows you to adjust the calculation timeframe, analyzing historical metadata from up to two years prior and forecasting up to three years into the future. Default alert policies have been established for supported devices based on the following condition:

  - Storage System Used Capacity >= 60 % In next 180 Days, generates Informational capacity alerts.
  - Storage System Used Capacity >= 75 % In next 90 Days, generates Warning capacity alerts.
  - Storage System Used Capacity >= 80 % In next 30 Days, generates Critical capacity alerts.
- Past date

  Specify how far into the past to use metadata for the capacity projection. In general, the further in the past that you select, the more metadata is used in the calculations, which can help aid the capacity projection.

  By default, the past date is 1 year before today's date. If metadata is unavailable at the selected past date, metadata is used from the date when it is first available. The minimum past date is 7 days before today's date, and the maximum past date is 2 years before today's date.
- Future date

  Specify how far into the future you want to project capacity usage. By default, the future date is 1 year from today's date.

For more information, see Capacity planning.

1.3 Notifying users before capacity runs out of space

The IBM Storage Insights capacity planning feature can send alerts in advance if any storage system touches the pre-set capacity limit or will run out of space shortly. This alert can be set through Alert Policies.

1. To view or create an Alerts Policies window, select **Configuration** → **Alert Policies**, as shown in Figure 1-3.

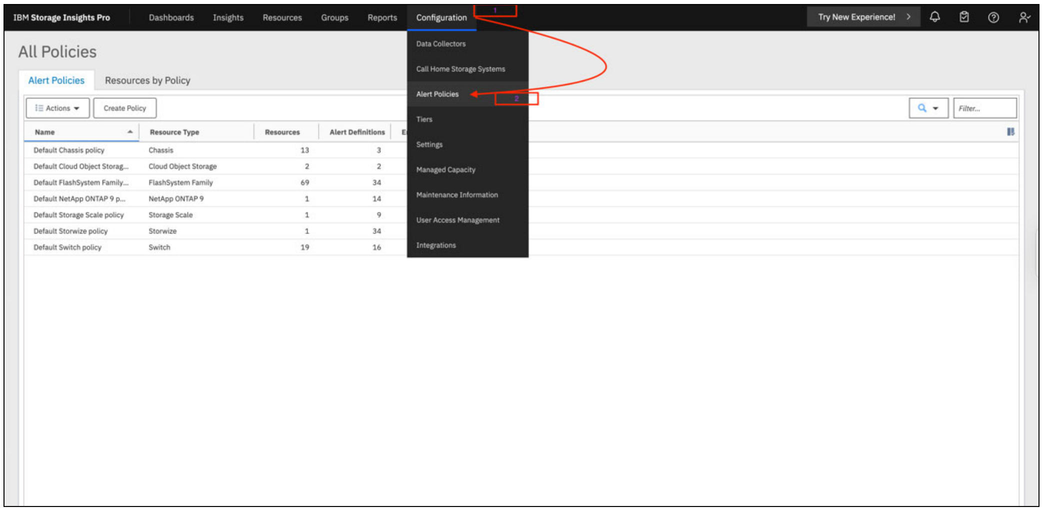


Figure 1-3 Alert Policies

For more information about creating an Alert Policy, see Alert policies.

2. To define a customized definition for Capacity alerting when you create an alert policy, select **Alert Definitions** → **Storage System** → **Capacity**. You can also customize the alert policy after it is created by editing it, as shown in Figure 1-4.

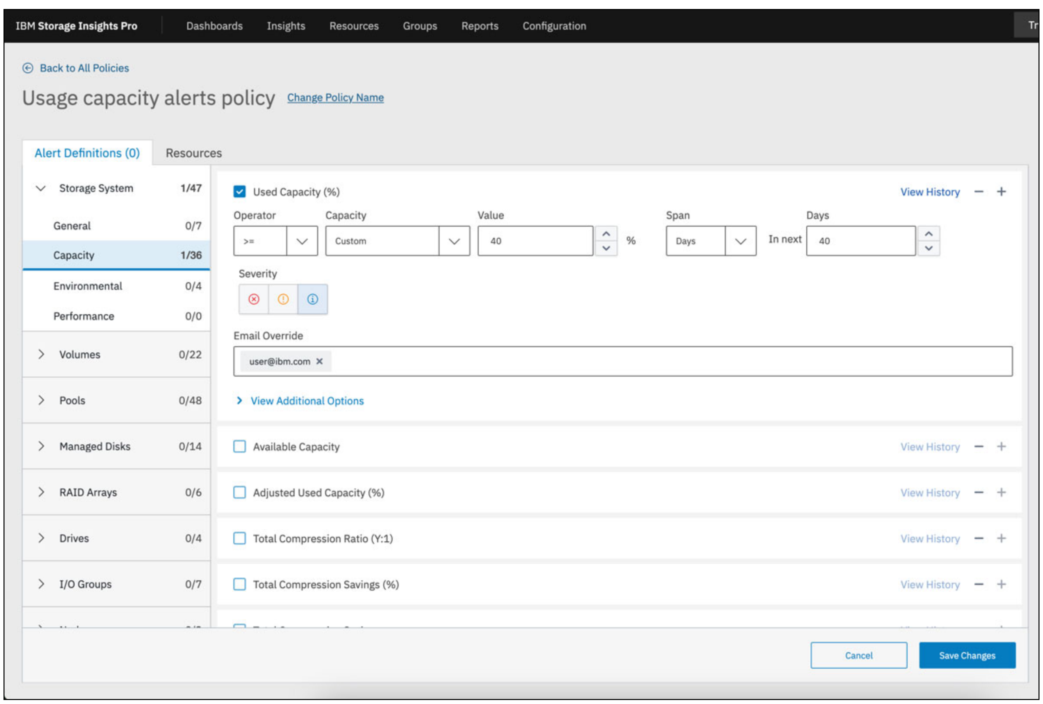


Figure 1-4 Editing an alert policy

3. Click the **Capacity** tab, and then select the **Used Capacity(%)** checkbox. Complete the following fields:
   - The Operator field specifies the conditions for generating an alert for an attribute. It has two options: >= or <=.
   - The Capacity field is provided with two drop-down values: Custom (input your wanted value in the Value field) and Capacity Limit (set Value to 80% by default).

- The Span field is used to specify the number of days in advance a user should be notified before a storage system reaches its capacity limit. This setting enables proactive capacity planning by ensuring that users receive alerts before storage runs out or reaches the predefined threshold.
- IBM Storage Insights leverages AI-driven algorithms to analyze historical storage usage patterns and predict when a system might reach critical capacity levels. Based on these predictions, the Span field helps users configure alerts in advance, giving them enough time to take necessary actions, such as optimizing storage, adding capacity, or reallocating resources.
- The Date field represents the predicted point in time when the forecasted capacity is expected to reach the defined capacity limit. It is used to trigger alerts, notifying users in advance so they can take necessary actions to manage storage resources efficiently. This setting helps with proactive capacity planning and prevents potential storage shortages or performance issues.
- The Severity field is where you select the severity of the alert.
- The Email Override field is used to specify the mail address to which email notifications are sent.

In this use case, the alert definition specifies that for a capacity attribute such as Used Capacity, the alert is generated 40 days in advance when the percentage of used capacity is greater than or equal to 40%, as shown in Figure 1-4 on page 6.

4. After completing the required details, click **Save Changes**. To see the alert policy, select **Configuration** → **Alert Policies**.

   For more information, see Defining alert definitions for general attributes and capacity changes.

   Now, if the Capacity planning feature predicts that the selected system in the alert policy definition will reach the pre-set value in pre-set days, then the user receives an email notification so that they can make changes or plan capacity requirements as needed.

1.4 Capacity reclamation

A key use case for the Capacity Reclamation feature in IBM Storage Insights is to proactively address and mitigate impending storage capacity shortages. By identifying and reclaiming unused or underutilized storage resources, this feature helps ensure business continuity and optimize storage resource utilization. For example, If capacity forecasting indicates that storage systems are likely to run out of space in the near future, customers can use the Capacity Reclamation feature to reclaim unused capacity. This proactive measure ensures business continuity by optimizing storage resource utilization.

To see the capacity reclamation overview window, complete the following steps:

1. Open the IBM Storage Insights instance dashboard, which is shown in Figure 1-5.

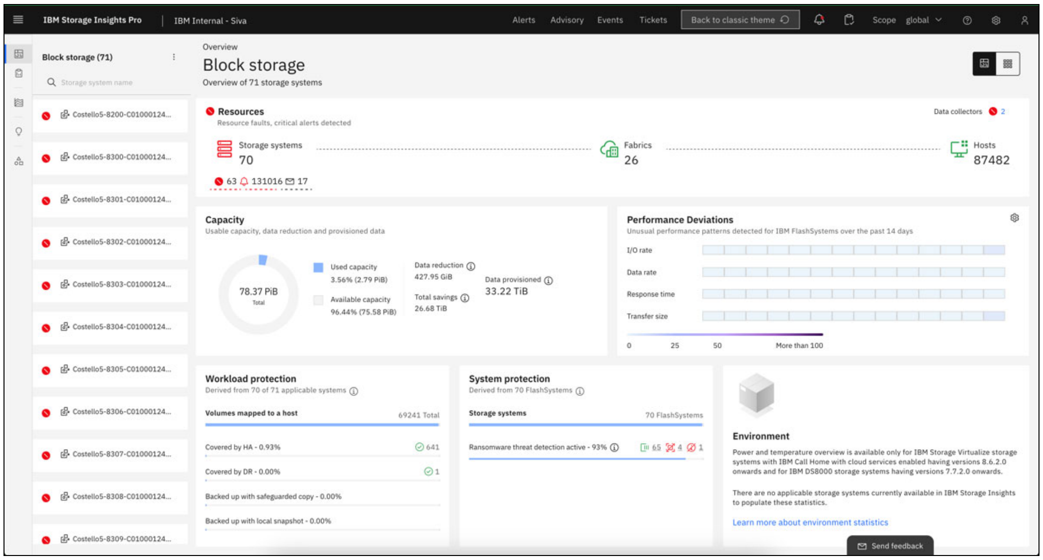


Figure 1-5   IBM Storage Insights instance dashboard

2. To open the Reclamation overview page, hover your cursor over the insights icon , click the insights drop-down icon, and click **Reclamation**.

Figure 1-6 shows the Reclamation overview window.

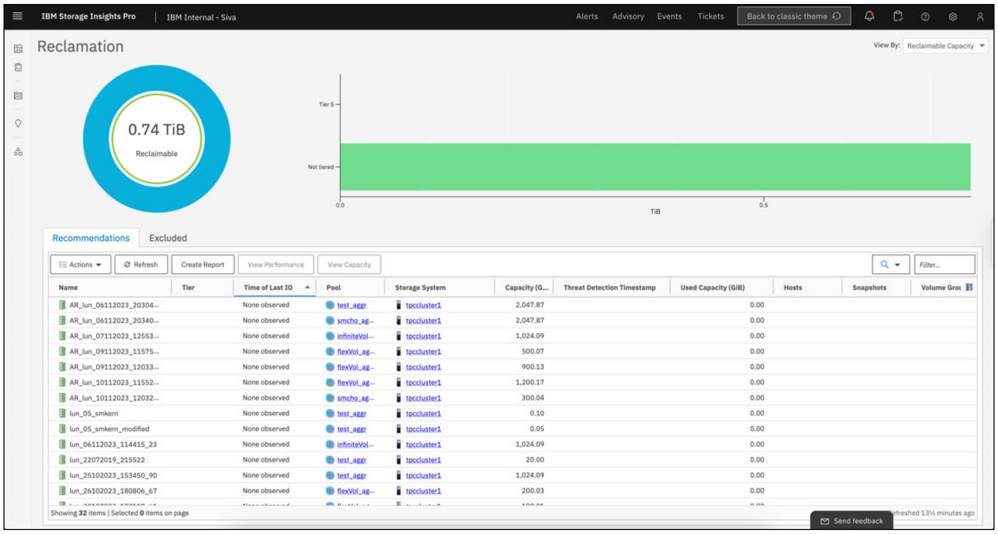


Figure 1-6 Reclamation overview page

Here is a breakdown of the reclamation donut chart in Figure 1-6.

- ► The circle in the upper left represents the amount of storage space that is used.
- ► The bar in the upper right represents the amount of storage space that can be reclaimed.
- ► The center of the donut chart provides a rounded estimate of the storage space that can be saved when the volumes that are listed in the table are reclaimed.

As shown in Figure 1-6 under the **Recommendations** tab, the process uses analytics to determine which volumes are eligible for reclamation. It evaluates storage capacity usage patterns and identifies underutilized volumes based on predefined criteria. These insights help in selecting potential candidates for reclamation, ensuring efficient storage management.

Information about reclaimable volumes

The following information is provided about each volume that is identified as a potential candidate for reclamation:

- ► Applications

  If the volume is added to an application, the name of the application is shown. If the volume is added to an application and its subcomponent, a number is shown. You can click the name or the number to view detailed information about the application.

- ► Available Capacity (GiB) (previously known as Unallocated Space)

  The total amount of remaining space that can be used by the volume, that is, the capacity that is not used by thin-provisioned volumes. This value is determined by the formula `Capacity - Used Capacity`.

  This value is not available for IBM FlashSystem A9000, IBM FlashSystem A9000R, and volumes from SpecV data reduction pools.

- ► Capacity (GiB)

  The total amount of storage space that is committed to a volume. For thin-provisioned volumes, this value represents the provisioned capacity of the volume.

- ► Hosts

  The name of the host to which a volume is assigned.

- Tier

  The tier level of the pool in which the volume is located. If the tier level is not defined for the pool in the data center, no value is shown.

- Time of Last I/O

  The number of days, weeks, or months when the value for the I/O rate of the volume was detected. The I/O rate is one of the determining factors for identifying the volumes that are candidates for reclamation.

- Used Capacity (GiB) (previously known as Used Pool Space)

  The amount of usable capacity that is taken up by data in a storage system after data reduction techniques are applied.

To see whether any volume is excluded from the recommendation, click the **Exclude** tab. Right click on the volume and select the option **Exclude from Analysis**.

For more information, see Reclamation views.

1.5 Recommendations if capacity reclamation does not work

The capacity reclamation feature of IBM Storage Insights can suggest capacity reclamation based on volumes or storage systems. Sometimes, the customer may not be able to reclaim the volumes because they are still needed for operations. In such situations, the organization can plan for extra storage capacity by considering new storage systems from IBM. They can also contact their IBM sales representative.

For more information about the IBM Storage systems, see More in store with the IBM Storage Platform.

2

Finding a suitable storage system for a new workload

As enterprise workloads become more dynamic and storage infrastructure becomes increasingly heterogeneous, the ability to place new workloads on the most appropriate storage system is critical. Effective placement ensures optimal resource utilization, maintains service level objectives (SLOs), and improves the operational efficiency of storage administrators. IBM Storage Insights introduces a purpose-built capability called *Workload Placement Advisor*, specifically designed to assess existing IBM Storage FlashSystem devices and determine the most compatible system for hosting new workloads.

The Workload Placement Advisor leverages historical system telemetry, performance forecasting models, and benchmark analytics to score and rank candidate systems. This chapter details the architectural design, core components, algorithmic logic, and user interface of this feature, offering a deep dive into how IBM Storage Insights simplifies workload placement using artificial intelligence and analytics.

This chapter has the following sections:

2.1 Workload placement planning considerations

Before initiating the selection of a storage system to host a new workload, it is critical to perform a thorough assessment of both the workload's requirements and the current capabilities of available storage systems. This step forms the foundation for accurate placement decisions, ensuring long-term performance stability and optimal resource utilization. The assessment involves multiple key dimensions, each of which contributes to the overall compatibility analysis:

2.1.1 Capacity and growth planning

The storage requirements of the new workload must be clearly defined in terms of initial capacity, as well as anticipated future growth. This includes both the primary data footprint and additional capacity overhead required for point-in-time copies, such as snapshots and safeguarded copies. For example, workloads that rely on frequent data protection or backup operations may introduce significant overhead, which must be accounted for during planning. Estimating growth over a defined period typically one to three years is essential for evaluating long-term sustainability on the target system.

2.1.2 Performance requirements and I/O profiling

Each workload introduces unique performance demands on the storage system. Key performance parameters to be specified include the total Input/Output Operations Per Second (IOPS), average read and write transfer sizes (typically measured in kilobytes per operation), and the expected read-to-write ratio. Furthermore, the expected end to end latency; the time taken to complete a read or write request is a critical constraint. A mismatch between these requirements and the storage system's capability can lead to performance degradation not only for the new workload but also for existing applications sharing the same infrastructure.

2.1.3 Workload duration and lifecycle

Understanding the intended duration of the workload whether it is short-term, seasonal, or long-term is another crucial factor. Temporary workloads (e.g., those aligned with promotional events or seasonal analytics) may be suitable for systems that have short-term capacity or performance headroom but would not support sustained growth. In contrast, persistent workloads require a longer forecasting window to ensure that resources remain sufficient throughout the workload lifecycle. This distinction directly impacts forecasting models and the selection of candidate systems.

2.1.4 Impact on existing baselines and resource forecasting

Introducing a new workload can alter the performance baseline of a storage system. It is essential to project the incremental resource consumption across capacity, IOPS, CPU utilization, data throughput, and latency, and evaluate whether these increases would push the system beyond acceptable operating limits. This is achieved through forecast modelling that combines historical telemetry with predictive analytics, allowing IT administrators to visualize the system’s projected state with the new workload included.

2.1.5 Compatibility, thresholds and operational constraints

Even if a system appears to have sufficient capacity or performance margin, it may not be suitable due to other system-level constraints. These include:

- CPU utilization nearing critical thresholds (typically above 60%–80%).
- Latency sensitivity for existing latency-critical applications.
- Pool configurations that complicate provisioning (for example, multiple tiered or deduplicated pools).
- Hardware limits such as maximum IOPS or bandwidth achievable for the given hardware and firmware version.

In addition to hard constraints, soft constraints such as administrative preferences may also influence system eligibility. The workload placement engine evaluates these factors holistically, ensuring that the selected system does not merely fit, but fits in a way that is sustainable, resilient, and aligned with policy-driven constraints.

2.2 Overview of the Workload Placement Advisor

The Workload Placement Advisor is an AI-driven feature available in IBM Storage Insights for IBM Storage FlashSystem customers. It performs the following steps:

1. Captures new workload specifications from the user.
2. Allows selection from predefined workload templates or entry of custom workload parameters.
3. Applies historical performance and capacity forecasting.
4. Invokes benchmark correlation through a performance library.
5. Scores and ranks candidate systems based on multidimensional fitment criteria.
6. Presents results with detailed visualizations and export options.

This functionality is accessible through the Carbon UI experience in IBM Storage Insights.

2.3 User interface and workflow

A workload is defined either by user-provided performance and capacity requirements or by selecting a pre-defined industry standard template, including read/write ratios, IO sizes, expected latency, and storage duration. IBM Storage Insights enables users to input these workload characteristics into the Workload Placement Advisor. Using AI-driven forecasting and fitment scoring logic, it evaluates the fleet of IBM Storage FlashSystem devices and ranks them by suitability.

The following are the steps for finding a suitable storage system for a new workload with Workload Placement Advisor:

1. Launch Workload Placement Advisor: Navigate to **Planning** → **Workload Placement Advisor** from the IBM Storage Insights UI.
2. Input workload parameters: Fill in capacity, yearly growth, read/write transfer sizes, IO rate, and response time expectations.

3. Modify forecast timeline: Choose custom past and future periods to improve prediction accuracy based on known usage.
4. Select storage systems for evaluation: Filter devices based on geography, FlashSystem grid membership, or business constraints.
5. Run evaluation: Trigger analysis. The system returns ranked systems with compatibility scores from 0 to 5.

2.3.1 Accessing the Workload Placement Advisor

Access the Menu from the Overview dashboard in IBM Storage Insights. Click **Planning** and then **Workload Placement**, as shown in Figure 2-1.

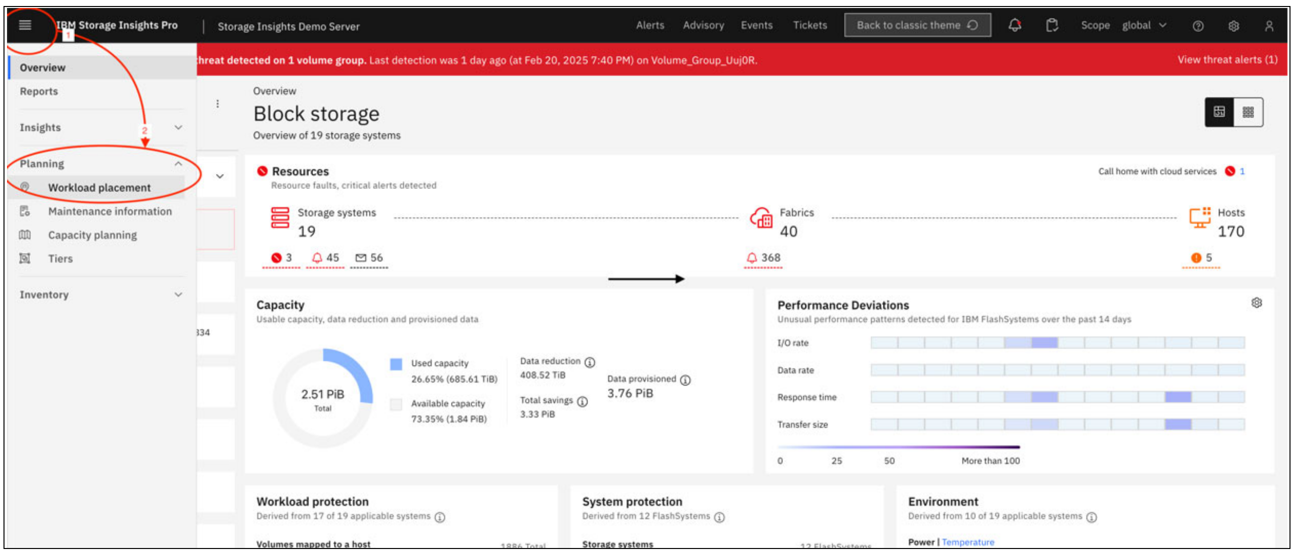


Figure 2-1 Launch Workload Placement Advisor

2.3.2 Workload definition and input

The Workload Placement Advisor supports:

- **Predefined templates**: OLTP, OLAP, Backup Appliance.
- **Custom workloads**: User-specified IOPS, transfer size, capacity, and so forth.
- **Input fields**:
  - Capacity (GiB)
  - Expected Yearly Growth (%)
  - Snapshot/Safeguarded Copy Switch
  - Total IOPS (ops/s)
  - Read/Write Ratio
  - Transfer Sizes (KiB/op)
  - Latency (ms)

Users can choose a t-shirt size (Small, Medium, Large) for predefined templates, or enter all values manually in the Custom template. See Figure 2-2.

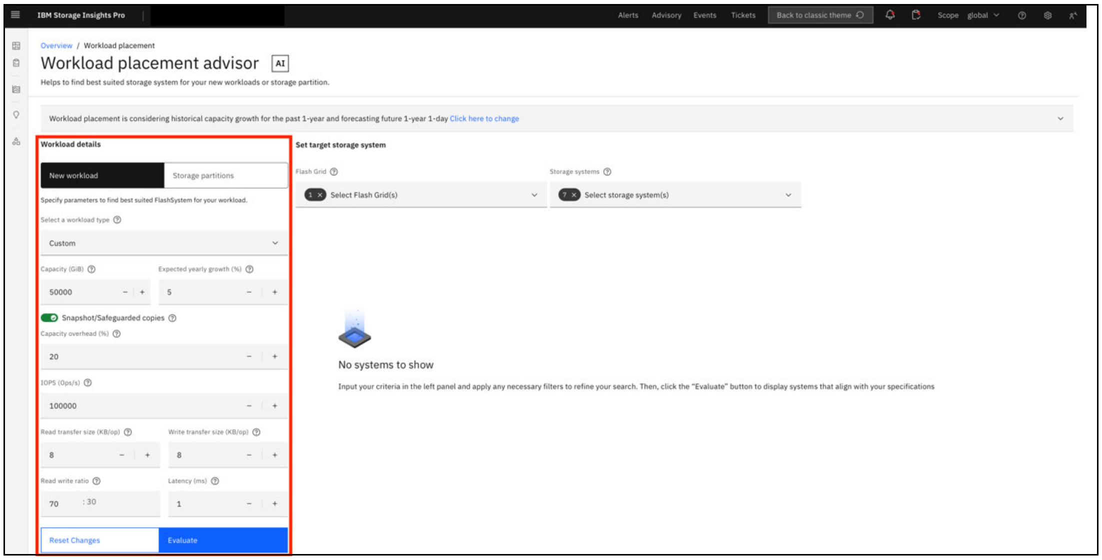


Figure 2-2   Figure 2.4.2-1 Specify the workload details

2.3.3 Forecasting time frame

Users can define:

- **Historical Range**: Up to 2 years in the past (primarily used for capacity data analysis).
- **Forecast Window**: Up to 3 years in the future (applies to both capacity and performance projections).

These settings are crucial for adapting the forecasting models to temporary workloads or recently re-purposed storage systems. The modification of the past and future dates may be required in some scenarios such as below:

- A storage system may be moved into production in the last three months and was earlier used for some other purpose such as a standby. In such cases changing the past date to reflect the three months of production can provide better results.
- The new workload being introduced may be a temporary workload to meet the business requirements such as the peak business season. In such cases, changing the future date to reflect the short-term requirement can help to find a storage system which is otherwise not suitable for a longer duration.

The default values if not modified are one year in the past for historical analysis and one year and one day in the future for the forecast analysis.

Post modification a user should click the **Apply** button to recalculate forecasts and update placement recommendations accordingly. See Figure 2-3 on page 16.

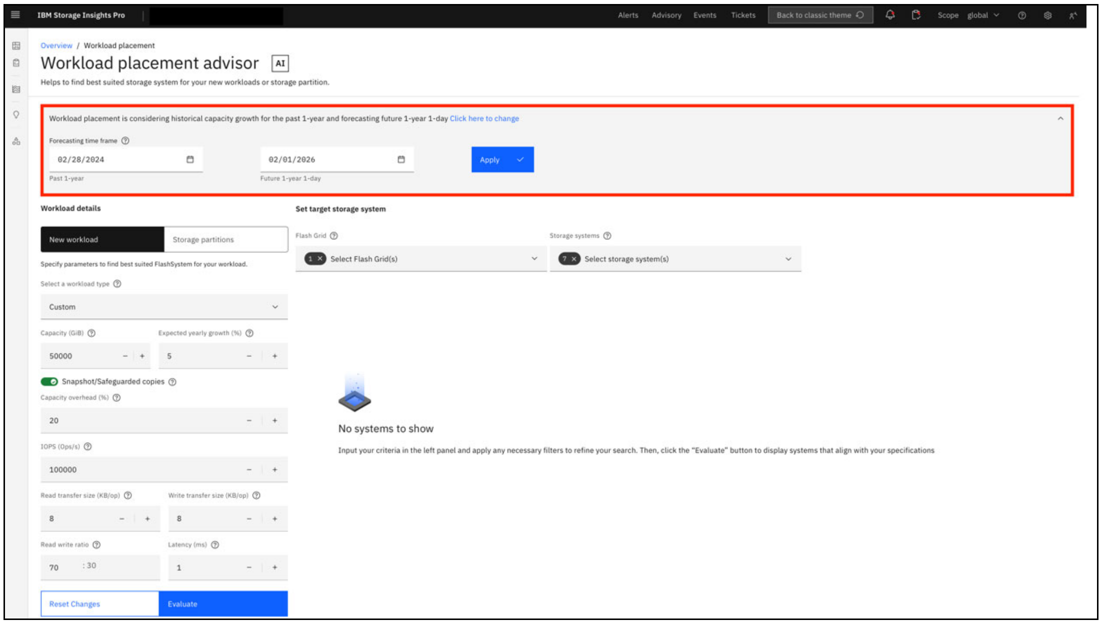


Figure 2-3 Modify the historical and forecasting time frame

System selection and filtering

Before initiating evaluation, the user can filter eligible systems using:

- ► Scope
- ► Flash
- ► System grid membership
- ► Storage system selector (multi-select dropdown)

By default, all the eligible storage systems available in the fleet are selected for the analysis. However, one should review and modify this list as per the business requirements. For example, the selected list of storage systems may include both the production and the development storage systems across multiple locations. Depending on the requirements of the new workload it may not be useful to do the analysis on all the storage systems and therefore the list should be filtered. This will help to evaluate the desired storage systems and will also help to reduce the time taken for the analysis.

The target storage systems can be filtered using the two available filters:

- ► FlashSystem grid: This filter allows you to select the storage systems which are part of a FlashSystem grid. If no FlashSystem grid exists, this filter will be empty. The storage system in the second filter will be modified based on the selection here.
- ► Storage systems: This filter allows you to select the desired storage systems.

See Figure 2-4 on page 17.

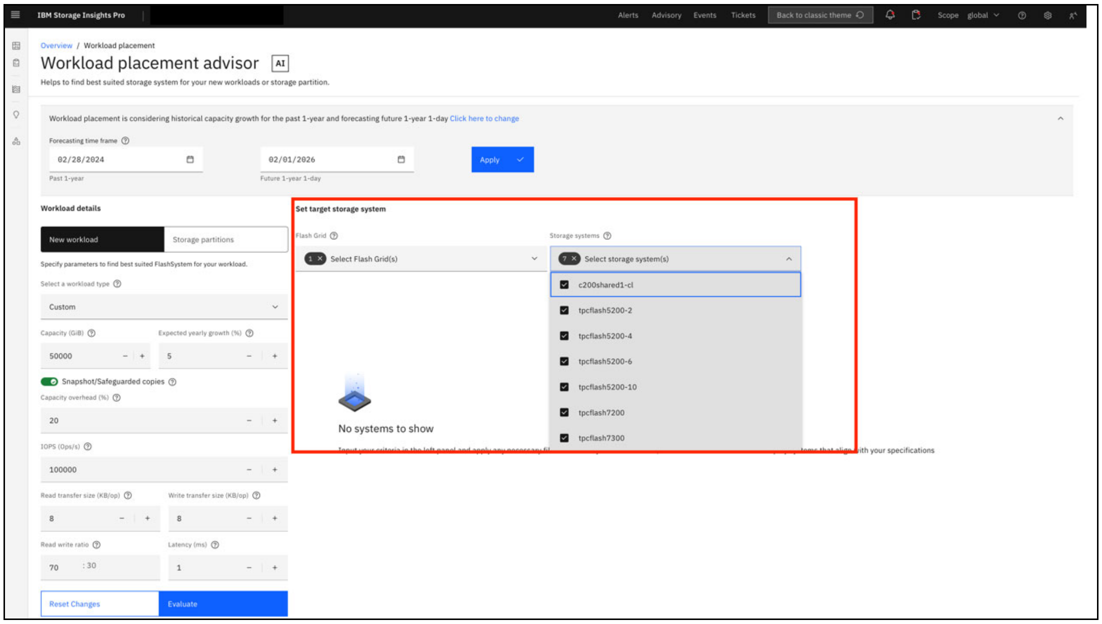


Figure 2-4 Set storage system search scope

Evaluation and results

Clicking **Evaluate** triggers the appropriate backend API to compute compatibility scores for each candidate system. Results appear on the right panel with a ranked list of systems.

After evaluation, IBM Storage Insights presents:

- **Compatibility Score**: Overall suitability of a device (0–5).
- **Forecast Charts**: Projected impact of the new workload on capacity, IOPS, data rate, latency, and CPU utilization.
- **Score Breakdown**: Visualization of how each metric (e.g., CPU utilization) contributes to the score.

The compatibility analysis checks whether introducing the new workload will exceed critical thresholds in any resource dimension. If yes, the system is marked as unfit (score = 0). Otherwise, a higher score indicates better fitment.

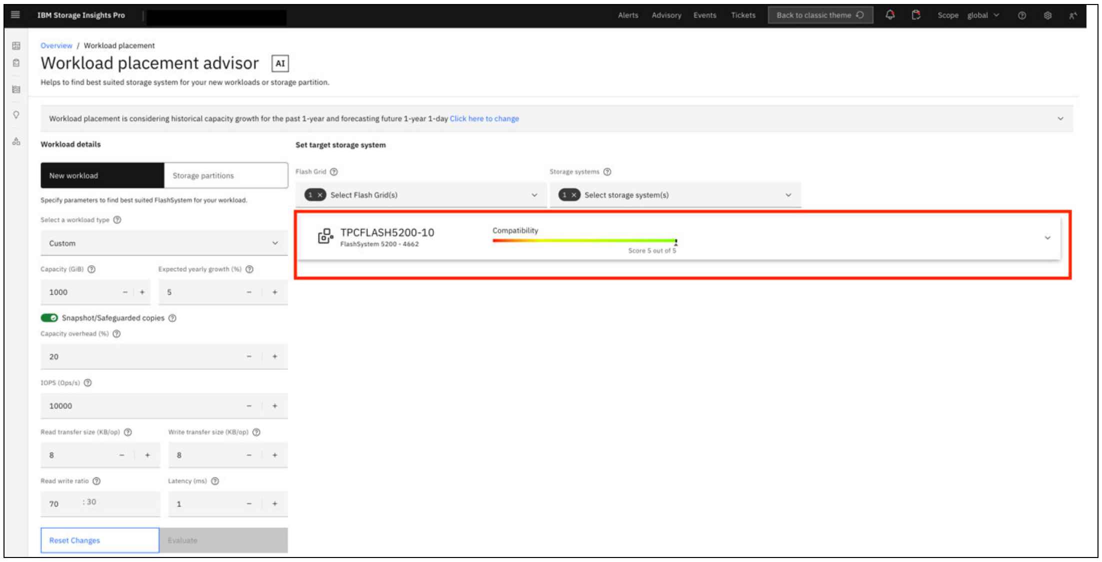


Figure 2-5 Review results

The results panel when expanded displays:

- Capacity growth chart before/after workload.
- Latency, IOPS.
- Data rate and CPU utilization forecasts with fitment scores.

A complete result card for one device showing metric-by-metric evaluation. See Figure 2-6/

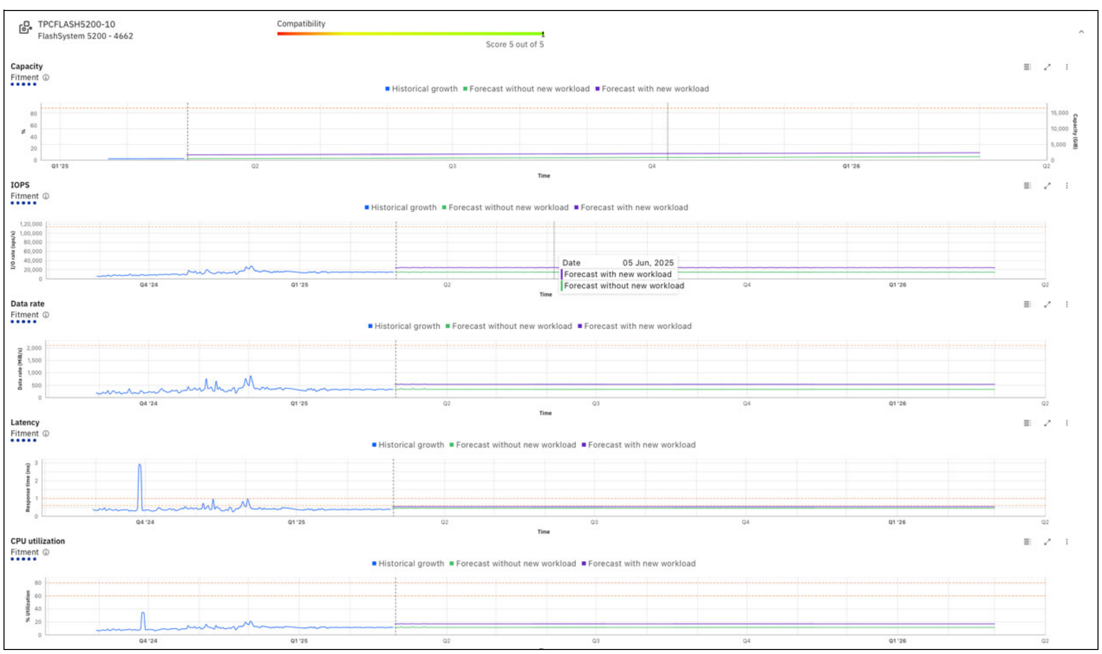


Figure 2-6 Detailed analysis of the result

2.4 Architecture and technical components

The Workload Placement Advisor is built on a distributed, modular architecture that brings together multiple subsystems to deliver a cohesive analysis platform. The design emphasizes scalability, extensibility, and robustness to support dynamic data inputs and real-time fitment evaluation all across the enterprise grade IBM Storage FlashSystem deployments.

2.4.1 Modular service architecture

At the core of the workload placement engine is a collection of independent services, each responsible for a distinct function. These services interact over secure internal APIs and process telemetry, forecast trends, and compute resource fitment in parallel to ensure timely and reliable recommendations.

Key components are as follows:

- **User Interface Layer**: Captures workload characteristics via a guided input form, offering users the choice between predefined templates and customizable workload profiles. The UI also manages data validation, state persistence, and triggers the evaluation workflow.
- **Input Processing and Configuration Engine**: Interprets the workload parameters provided by the user, including capacity, IOPS, latency, read/write ratio, and data transfer sizes. For predefined templates, it maps selected workload types (e.g., OLTP, OLAP, backup) to internally defined benchmarks.
- **Forecasting Engine**: Consumes historical storage system telemetry to project future trends in capacity, CPU utilization, IOPS, latency, and bandwidth. Forecasts are metric-specific and tuned to the operational history of each system, ensuring high relevance and precision.
- **Benchmark Correlation Module**: Uses a performance benchmark library that encodes device-specific characteristics. This module estimates how a storage system would respond to additional workload by applying empirical performance curves based on device type, model, and workload profile.
- **Fitment Evaluation Engine**: Integrates forecasted trends, benchmark outputs, and user constraints to compute a multi-dimensional compatibility score. This scoring accounts for both absolute thresholds and relative system behavior under projected load.
- **Result Orchestration and Presentation**: Combines fitment scores and system metadata to create a user-friendly, ranked list of recommendations. Detailed visualizations allow administrators to explore each candidate system's forecasted performance across relevant metrics.

This loosely coupled design allows individual components to evolve independently, supporting new forecasting techniques, additional workload templates, or enhanced telemetry ingestion without requiring changes across the stack.

2.5 Fitment algorithm and scoring logic

The *fitment scoring algorithm* is central to the workload placement process. It evaluates how well a given storage system can accommodate a new workload without degrading performance or violating operational constraints.

The scoring model is both *metric-driven*, and *threshold-aware*, applying both hard and soft constraints to ensure comprehensive validation.

2.5.1 The core metrics that are evaluated

The core metrics that are evaluated are as follows:

- Capacity Fitment
  - Assesses the projected capacity usage of the new workload.
  - Includes overhead for snapshots or safeguarded copies if enabled.
  - Compares forecasted year-end usage to available usable capacity.
  - Surplus capacity determines score (0–5) based on percentage thresholds.
- IOPS
  - Measures total projected IOPS with the new workload added.
  - Evaluates proximity to device-specific IOPS saturation point (as derived from the benchmark library).
  - Scoring is tiered by surplus margins over safe operational thresholds.
- Latency
  - Forecasted latency is compared against user-defined maximum acceptable latency.
  - Latency penalty is applied if predicted values exceed user thresholds or current system baselines.
  - A two-tier comparison model is used:
    - Forecasted latency versus user target.
    - Forecasted latency versus average existing latency.
- CPU Utilization
  - Projects CPU usage with new workload.
  - Device-specific CPU correlation is derived from benchmarking data.
  - Score reduced or eliminated if CPU crosses soft (60%) or hard (80%) thresholds.
- Bandwidth (Data Rate)
  - Treated as a function of IOPS and transfer size.
  - Benchmarked data provides expected throughput limits for different workloads.
  - Shares score with IOPS for correlation consistency.

2.5.2 Score aggregation

Each metric is assigned a normalized score between 0 and 5. A score of zero in any single metric signifies incompatibility, resulting in an overall score of zero. The final compatibility score is determined through a weighted approach.

Composite Score Formula

Let us assume:

- s_c, s_i, s_{cpu}, s_l are scores for capacity, IOPS, CPU, and latency respectively.
- w_c, w_i, w_{cpu}, w_l are respective weights (default = 1).

Final Score = $(w_c \times s_c + w_i \times s_i + w_{cpu} \times s_{cpu} + w_l \times s_l) / (w_c + w_i + w_{cpu} + w_l)$.

2.5.3 Confidence modifiers and guardrails

Systems with multiple pools or data reduction pools may have score penalized to reflect configuration complexity. Systems where fitment assumptions (e.g., latency tolerance) break down due to edge conditions are adjusted accordingly. Compatibility labels (For example, Fully Suitable, Conditionally Suitable, Not Suitable) are derived from the final score.

2.6 Performance forecasting and data handling

Forecasting forms the analytical backbone piece of the workload placement engine. Rather than relying solely on current system state, forecasts allow predictive modelling of how storage systems will behave under increased load, ensuring that short-term suitability aligns with long-term sustainability. Each metric (Capacity, IOPS, CPU, Bandwidth, Latency) is forecasted independently using a tailored time series approach.

2.6.1 Forecasting workflow

IBM Storage Insights leverages advanced time series forecasting methodologies to estimate future values of key storage performance metrics, including IOPS, latency, data rate, and CPU utilization. These predictions are used to evaluate how introducing a new workload may impact the performance posture of a target storage system.

The forecasting pipeline consists of the following stages:

Data Ingestion

Telemetry data is collected at fine-grained intervals (e.g., 5-minute resolution) across a span of up to six months for each storage system. These metrics represent historical usage patterns and serve as the basis for forecasting. Data is grouped by storage system and logical components, such as partitions.

Data Resampling

To support robust long-term forecasting, high-frequency data is resampled to daily granularity. A 95th percentile filter is applied during resampling to capture upper-bound operational behavior while smoothing out anomalies. This ensures that the forecast reflects conservative estimates aligned with peak usage conditions.

Model Application

To generate reliable predictions for future system performance, IBM Storage Insights employs a hybrid forecasting approach that leverages both statistical modelling techniques and compact foundation models purpose built for time-series data.

- Classical Time Series Models: These models are optimized for storage-specific time series and learn patterns such as trend, seasonality, and autocorrelation. They form the first layer of analysis and provide a robust statistical foundation for forecasting.
- Tiny Time Mixers (TTMs): Developed by IBM Research, TTMs are a family of compact, pre-trained models designed specifically for multivariate time-series forecasting. TTMs are pre-trained on publicly available time-series datasets using various augmentations and are optimized for specific forecasting configurations. TTMs follow a focused pre-training strategy, where each model is tuned for a specific context and forecast length. This approach ensures high accuracy, minimal resource usage, and rapid inference; making it ideal for real-time advisory tasks in storage environments.

This model ensemble supports robust forecasting for IOPS, latency, data rate, and CPU utilization across diverse workload conditions. By combining statistical precision with deep-learning adaptability, IBM Storage Insights ensures reliable resource planning and storage system recommendations.

Forecast Generation

Using the trained models, future values are projected across configurable time windows, typically spanning one year. These forecasts are then used to compare current system performance baselines with predicted metrics under the influence of the new workload. The result supports decision making during workload placement analysis, providing visibility into capacity constraints, potential bottlenecks, and long-term sustainability.

2.6.2 Forecast integration with benchmarking

Forecasts provide projected workload behavior, while the benchmark correlation engine translates these into expected system responses. For instance:

Forecasted IOPS is passed to the benchmark module to compute corresponding CPU load and latency.

Forecasted capacity is evaluated against system limits with growth margin estimations.

This two-step approach ensures that forecasted load is interpreted in the context of hardware-specific behavior, not just generic resource usage.

2.6.3 Best practices for finding a suitable storage system

The best practices for finding a suitable storage system for a new workload are listed below:

- Ensure at least 6 months of telemetry data for target systems.
- Regularly review benchmark for hardware changes.
- Maintain up to date workload templates for frequently deployed applications.
- Use historical/forecast tuning controls for specialized workload durations.

2.7 Conclusion

IBM Storage Insights offers the Workload Placement Advisor, a data-driven, AI-enhanced decision support system for storage administrators. Its integration of forecasting, benchmarking, and optimization streamlines the planning process and guarantees workload deployment on the most suitable storage infrastructure.

3

Detecting performance deviations

Performance analysis is a critical aspect of debugging workload issues such as high latency and low throughput. Customers often face challenges in identifying performance deviations from their baseline and correlating them with configuration changes to pinpoint the root cause. For instance, a storage system upgrade could be the source of performance problems.

IBM Storage Insights gathers, processes, and visually presents extensive performance metrics from storage systems. The Performance Deviation widget leverages intelligent models to analyze time series data for designated key performance indicators (KPIs), detecting and displaying deviations in performance metrics. A deviation refers to a sudden spike in a metrics value for a storage system. The following four metrics are monitored for performance deviations:

- I/O rate
- Data rate
- Response time
- Transfer size

The widget is exclusively supported for block storage systems and requires a minimum of 14 days of data to initiate performance deviation monitoring. It is available to both IBM Storage Insights and IBM Storage Insights Pro users, although the detailed deviation view is accessible only to IBM Storage Insights Pro users. See IBM Documentation for more information.

This chapter has the following sections:

- “Performance Deviation widget” on page 24
- “Performance Deviation widget: Extensive view” on page 24
- “Settings” on page 26

Note: Workload Anomaly Detection identifies unusual patterns, but Performance Deviation Monitoring detects sudden spikes in KPIs, making them distinct features. For more information, see Alerts for workload anomaly detection and Performance Deviation.

3.1 Performance Deviation widget

The Performance Deviation widget summarizes performance deviations for block storage systems, as shown in Figure 3-1. It visualizes deviations for each metric over the past 14 days, with each rectangular box representing a single day. The timeline progresses from 14 days ago to yesterday, moving left to right, and is based on the Coordinated Universal Time (UTC) time zone. Users can click a box to view deviation data for a specific metric on a selected day. The color of each box reflects the number of deviations that were recorded on that day, with a color bar legend positioned at the lower left of the widget. Clicking a box reveals a more detailed breakdown of deviations for that day, and hovering your cursor over a box displays a tooltip with additional details, such as the number of deviations that are detected for the corresponding metric.

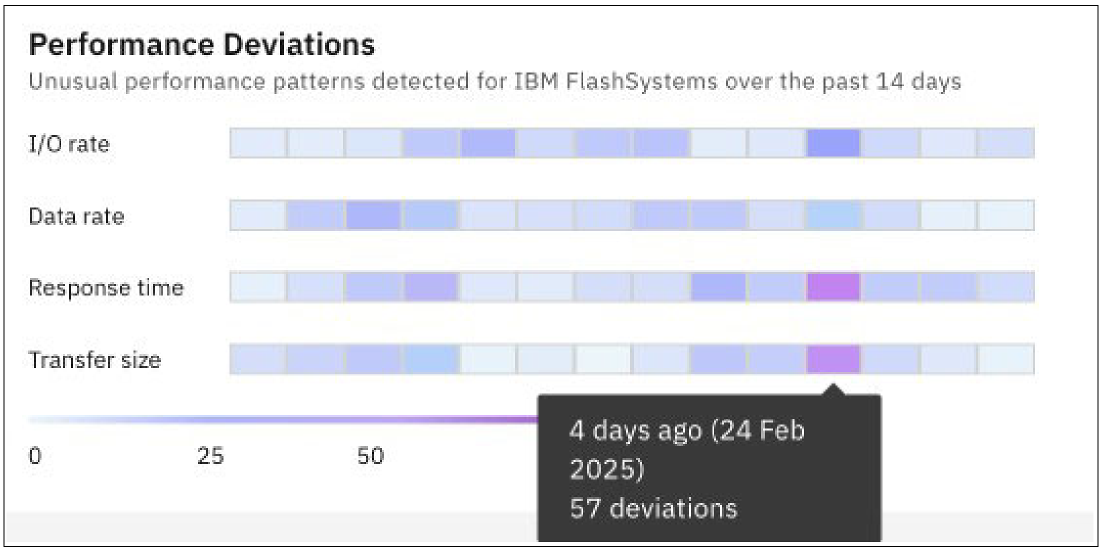


Figure 3-1 Performance Deviation widget overview

3.2 Performance Deviation widget: Extensive view

When you select a rectangular box for a specific metric in the widget, a side window appears and displays a list of storage systems that experienced performance deviations on the chosen day, as shown in Figure 3-2 on page 25. The deviations are presented in a tabular format, and categorized by storage system. Expanding a storage system reveals a more detailed breakdown of the deviations, as described in Table 3-1 on page 25. Also, users can download the deviation report for the selected day by clicking the download icon, as shown in Figure 3-3 on page 25.

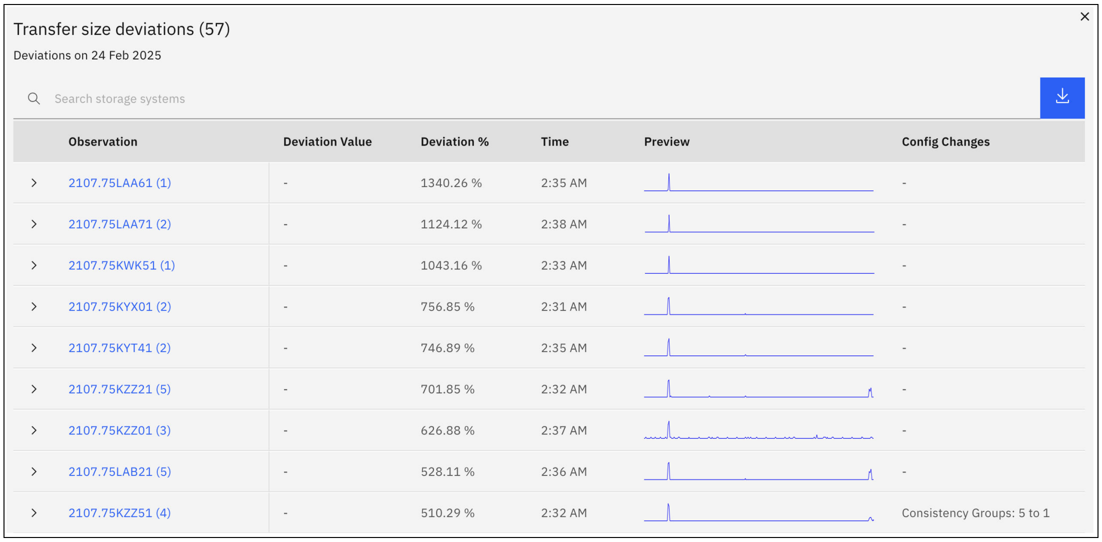


Figure 3-2 Deviations at system level

Table 3-1 Descriptions of the deviations

Label	Description
Observation	Represents an increased or decreased metric value.
Deviation Value	Indicates the metric value that was observed for the deviation.
Deviation %	Represents how much the observed value has deviated from threshold.
Time	Indicates the time at which the deviation occurred.
Preview	Shows a graphical representation of deviations that were detected in the last 24 hours.
Config Changes	Lists the configuration changes that happened in a storage system between the previous and current day.

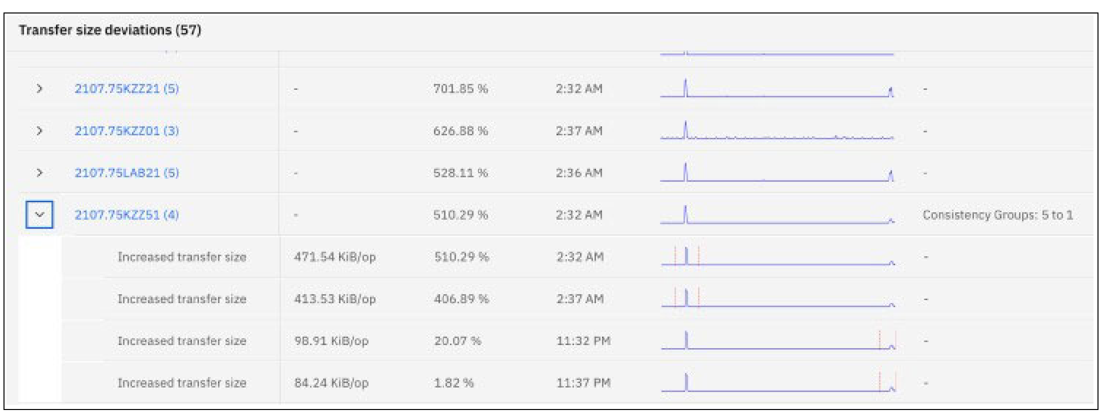


Figure 3-3 Granular level deviation details at the system level

3.3 Settings

Clicking the settings icon in the upper right of the widget opens the **Settings** dialog (Figure 3-4), where you can configure deviation thresholds and system exceptions, such as the following ones:

- Sensitivity: A slider lets users adjust sensitivity levels to Low, Medium (default), or High. A higher sensitivity detects more deviations.
- Exempted Systems: Users can exclude storage systems from monitoring by selecting them from a drop-down list, which displays all available systems with checkboxes. Selected systems appear under Exempted Systems on saving. To remove a system, click the delete (**x**) icon next to it.

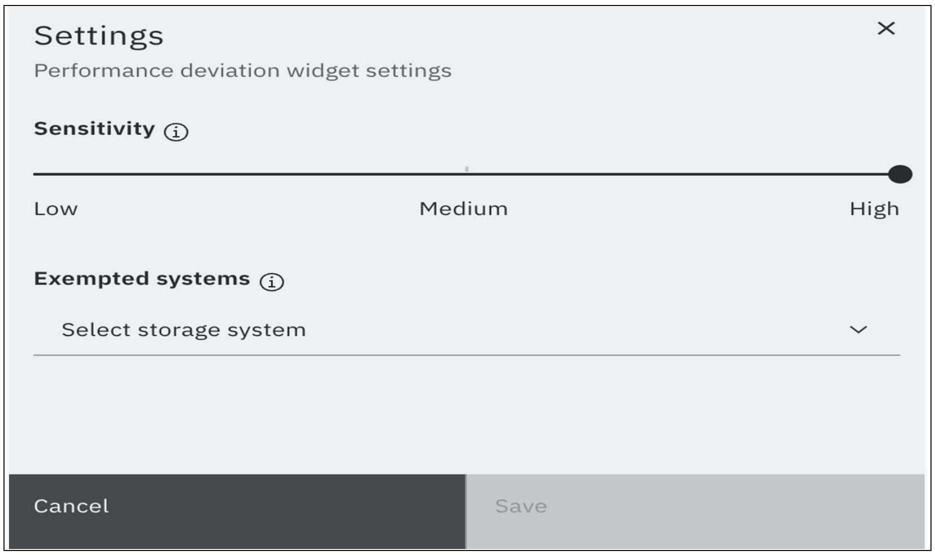


Figure 3-4 Settings widget to manipulate deviation parameters

Click **Save** to apply changes or **Cancel** to discard them.

For more information, see Performance Deviation.

4

AI-based Workload Advisory for storage partition

The AI-based Workload Advisory for storage partitions in IBM Storage Insights is a feature that leverages artificial intelligence to help storage administrators make informed decisions about managing and migrating storage partitions within an IBM FlashSystem grid.

Below are the sections contained within this chapter, which aims to provide a high-level understanding of the AI-based Workload Advisory for storage partitions:

4.1 Storage partition and FlashSystem grid

IBM FlashSystem is a high-performance, enterprise-grade storage platform designed to deliver ultra-low latency, scalability, and efficiency for modern workloads. One of the key architectural features of IBM FlashSystem is its ability to manage storage partitions, which provide secure and logical separation of storage resources within the system.

A *storage partition* in IBM FlashSystem refers to a logical boundary within the storage system that isolates resources, such as volumes, host mappings, and performance metrics, for different tenants, departments, or applications. This capability is essential in multitenant environments, managed service providers, and large enterprises where different teams or clients may share the same physical infrastructure but require isolated access and administrative control.

Each storage partition can be associated with specific *hosts and host clusters*, allowing administrators to define which systems can access the volumes within that partition. This logical segregation enhances security, enables role-based access control, and simplifies management by clearly delineating ownership and usage boundaries.

Storage partitions also enable performance isolation through per-partition QoS (Quality of Service) settings, ensuring consistent workload performance. By leveraging storage partitions, organizations can optimize resource utilization, enforce access policies, and streamline operational workflows, positioning IBM FlashSystem as an ideal platform for complex, multi-application environments.

The *IBM FlashSystem grid* is a conceptual and functional abstraction used within the IBM FlashSystem ecosystem to manage and operate distributed storage systems in a unified and scalable manner. It represents a logical grouping of multiple FlashSystem storage systems—physical or virtual—that are managed as a single cohesive entity within IBM Storage Insights, IBM's cloud-based monitoring and management platform.

A FlashSystem grid aggregates multiple IBM FlashSystem instances (such as FlashSystem 5200, 7300, and 9500), often deployed across different data centers or within a large-scale environment. By organizing these systems into a grid, IBM enables centralized monitoring, workload placement advisory, performance analytics, and health management across a federated storage infrastructure.

The grid model simplifies storage administration by abstracting the complexities of individual systems and presenting a unified interface through which storage administrators can perform operations such as:

- Assessing performance trends across the fleet.
- Making workload placement decisions.
- Managing storage partitions.
- Tracking capacity growth and usage patterns.
- Identifying bottlenecks or anomalies across systems.

Leveraging historical data analysis and AI-powered advisories, IBM Storage Insights, in conjunction with the FlashSystem grid, offers users a comprehensive solution for planning, monitoring, and managing their storage environment within a FlashSystem grid.

For a detailed understanding of FlashSystem grid and storage partition, consult the IBM Documentation on FlashSystem grid.

4.2 Workload Placement Advisory for storage partition

A *workload* can be defined as any service, application, or capability that consumes resources, such as storage. For example, a database is a workload using computing and storage resources. IBM Storage Insights helps customers identify best-suited resources for their workloads by leveraging historical data to provide advice for identifying ideal systems for any new workload deployment.

IBM Storage Insights offers an AI-driven Workload placement advisor to simplify the identification of the optimal storage system for new workloads. This feature uses AI to forecast capacity and performance needs, calculating a compatibility score for each selected target storage system. This score, along with detailed performance metric illustrations, provides strong justification for selecting the best storage. The analysis considers metric thresholds when evaluating suitability and is exclusively available for IBM Storage FlashSystem.

IBM Storage Insights focuses on the following aspects to generate the AI-based advisory for workload placement

- **Capacity:** Define the workload’s initial capacity needs and its growth, which will be used to find the storage system capable of sustaining the initial requirement and growth in future.
- **Performance:** Like capacity needs, it’s important to define the performance parameters and evaluate the same for any placement advisory.
- **Time:** Time plays an important role in understanding the resource requirement and growth capability; it helps evaluate the systems best suited for the given time frame.

The following sections describe the detailed steps for evaluating the best-suited storage system for a new custom workload.

4.2.1 Starting the Workload Placement Advisor

To access the Workload Placement Advisor, go to the **Planning** section within the IBM Storage Insights left panel (accessible from any page) as shown in Figure 4-1 on page 30.

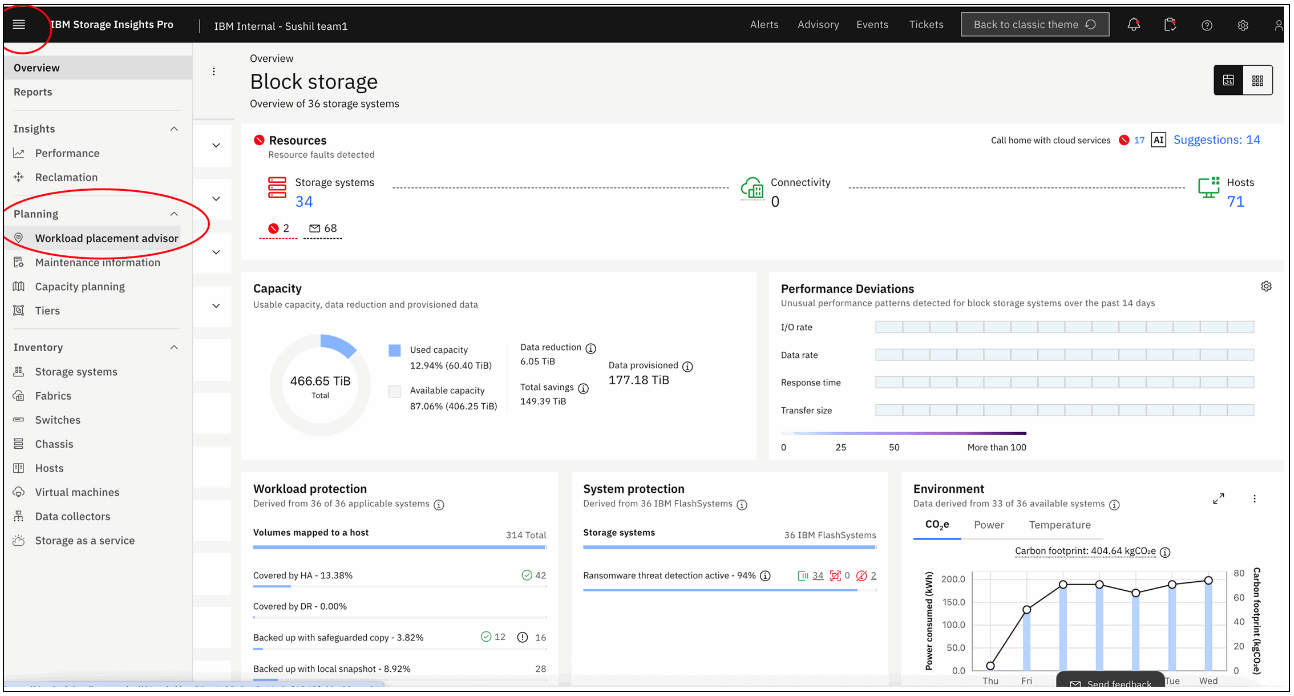


Figure 4-1 Workload placement

4.2.2 Specifying the workload details

Similar to other planning pages, the Workload Placement Advisory page allows users to begin by inputting their requirements. It collects information regarding initial and projected capacity growth, anticipated performance, and the duration for which the workload will be hosted. Users also have the option to select from predefined standard workloads such as OLTP, OLAP, and Backup Appliance. Choosing these will automatically populate the workload parameters based on industry benchmarks for Small, Medium, and Large deployments. Figure 4-2 illustrates the Workload Placement screen.

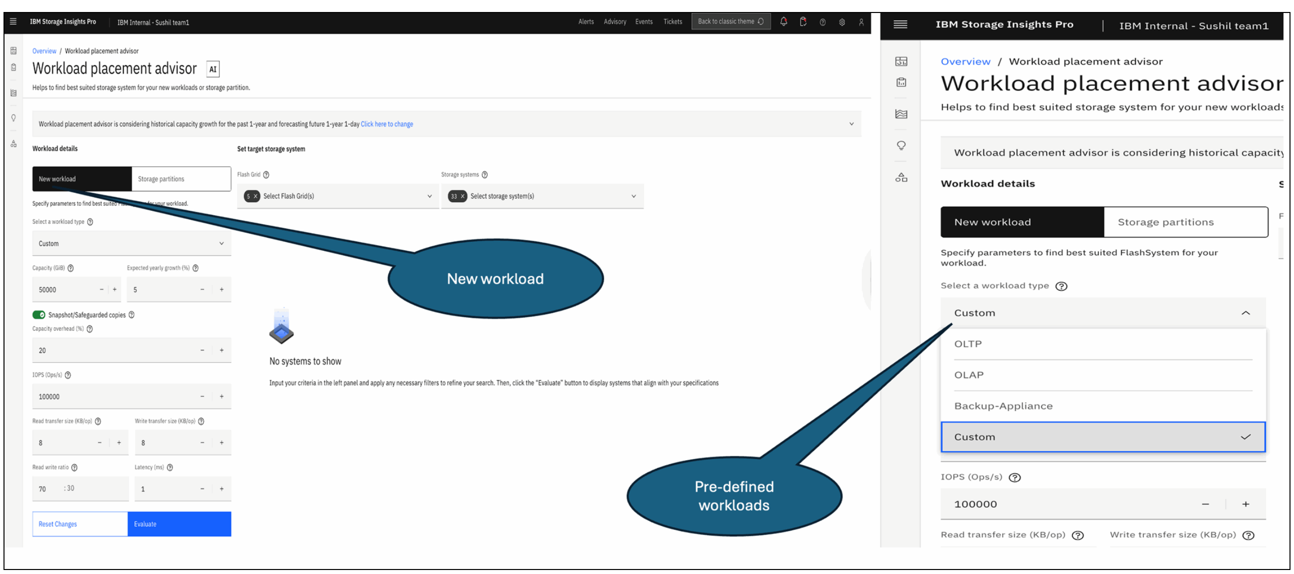


Figure 4-2 Workload Placement Advisory input parameters

4.2.3 Selecting storage partition

As described 4.1, “Storage partition and FlashSystem grid” on page 28, storage partition is a logical grouping of storage resources that represent a workload in storage system world. The workload placement advisory also has an option to run a placement for this storage partition to predict the best suited system to handle the partition growth and performance. Figure 4-3 illustrates the selection of Storage Partition on the Workload Placement Advisor page.

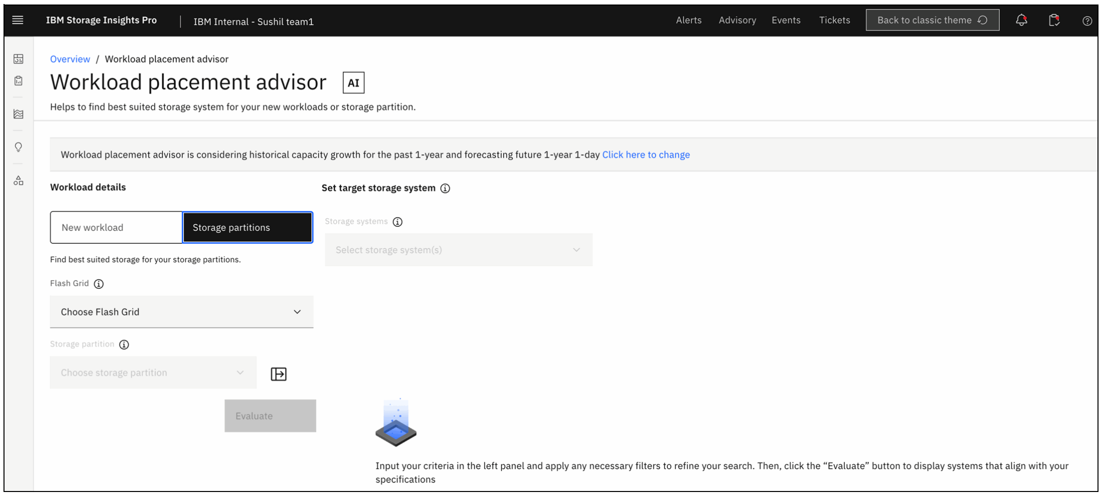


Figure 4-3 Storage partition – Workload Placement Advisor

When initiating a placement advisory, users can select the appropriate FlashSystem grid and choose the specific storage partition under evaluation, as illustrated in Figure 4-4 on page 32.

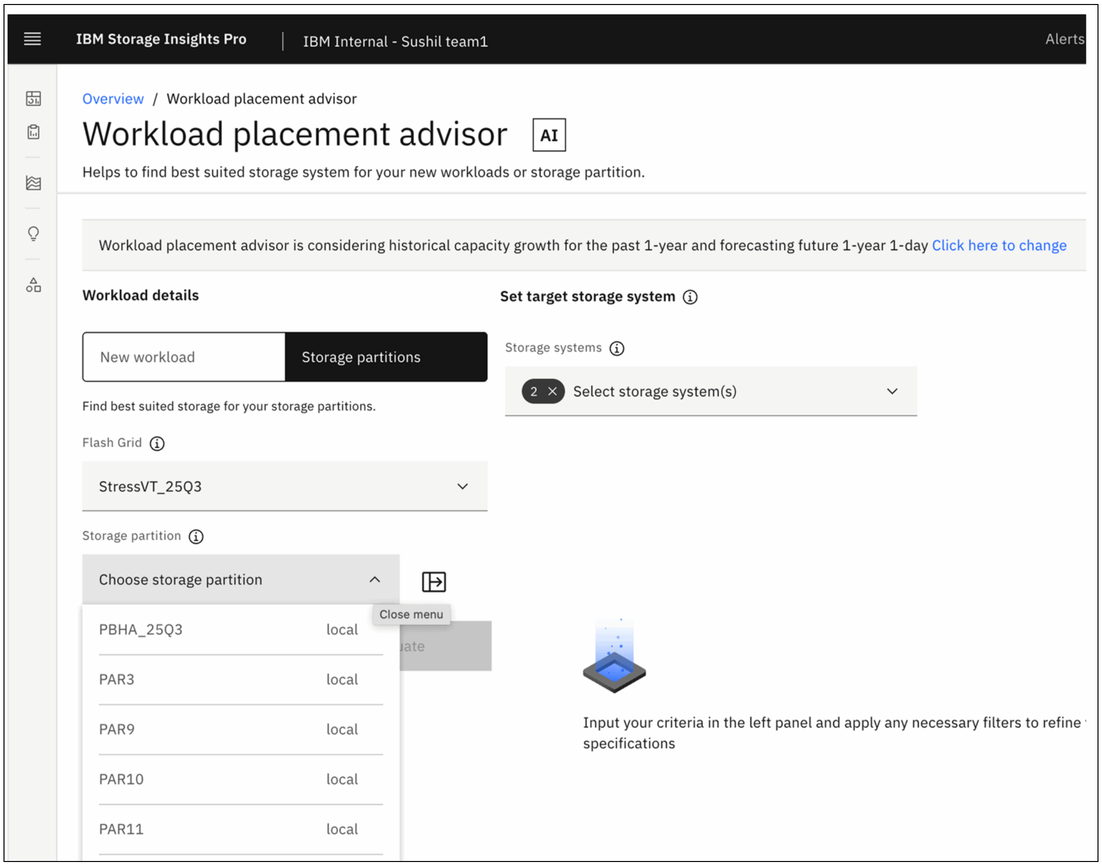


Figure 4-4   Select storage partition for Workload Placement Advisory

In IBM FlashSystem, a storage partition is designed to encapsulate a workload, eliminating the traditional need for users to manually define workload characteristics when placing new workloads. IBM Storage Insights enhances this process by using the partition's historical telemetry data, enabling a data-driven evaluation of trends in both capacity utilization and performance growth over time. To provide accurate and context-aware recommendations, the Workload Placement Advisor interface allows users to adjust the historical analysis period, ensuring that suggestions reflect relevant usage patterns and the workload's temporal dynamics, as shown in Figure 4-5. This integration streamlines decision-making while promoting consistency and efficiency in storage resource planning.

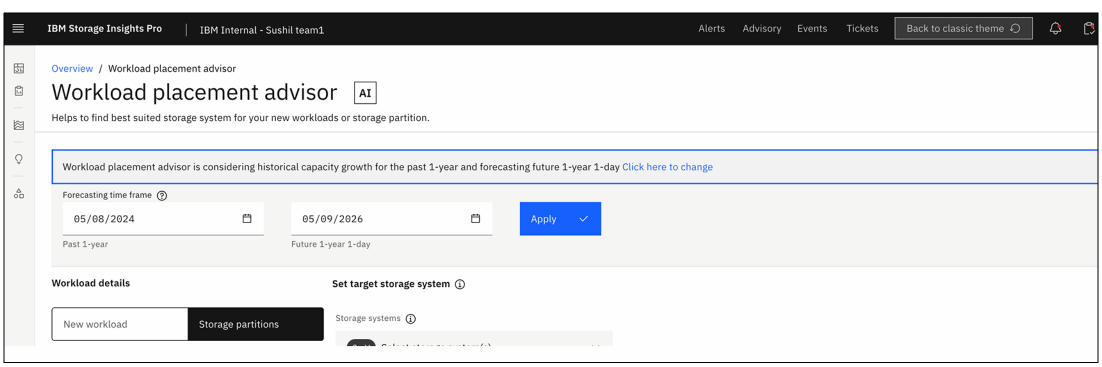


Figure 4-5   Adjust the timeline for historical and future for precise prediction

Note: Evaluation of storage partitions is restricted to those residing within the FlashSystem grid. As a result, only systems that are members of the FlashSystem grid constitute viable targets for storage partition mobility.

4.2.4 Understanding the Workload placement Advisor results

Leveraging the Workload Placement Advisor in IBM Storage Insights, users gain a detailed analysis for each storage system's partition. This includes a compatibility score alongside a visual forecast of capacity and performance utilization. Figure 4-6 exemplifies the placement advisory response for a storage partition. To gain a deeper understanding, let us examine the specifics.

Figure 4-6 Sample output of storage partition advisory

The sample in Figure 4-6 shows the storage system to be compatible with the new workload, with a compatibility score of 5 on a scale of 0 to 5. This overall score is calculated based on the scoring of the following metrics:

- **Capacity:** The capacity analysis considers the historical capacity utilization and the forecasted capacity growth with and without the storage partition. The detailed view of the capacity chart is illustrated below in Figure 4-7.

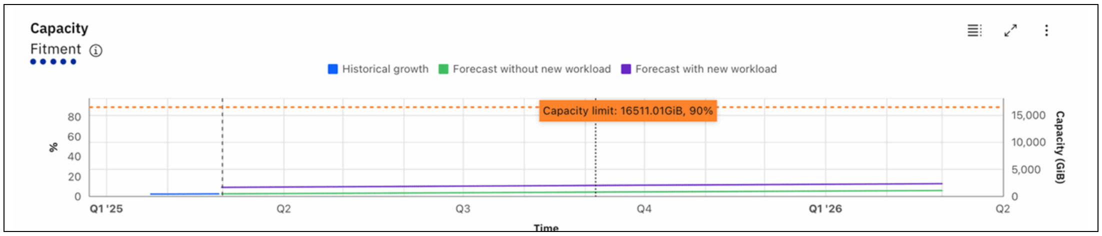


Figure 4-7 Capacity growth for storage partition

Figure 4-7 shows the new workload's forecasted capacity growth is under the 90% limit, resulting in a perfect capacity analysis score of 5. This user-defined limit defaults to 100% if not specified. Therefore, hosting the new workload on this storage system is deemed suitable..

- **IOPS:** IOPS represents the total I/O rate observed on the system and is measured in operations per second. The analysis considers the historical I/O rate and the observed growth rate to forecast the future I/O rate. The forecast is plotted to show the I/O rate with and without the new workload, as illustrated in Figure 4-8 .

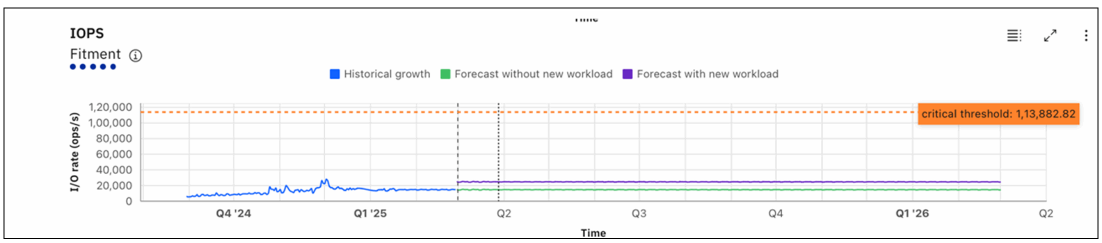


Figure 4-8 IOPS metric analysis

The expected IOPS on the target storage, as seen in Figure 4-8, are below the critical threshold determined by the Workload Placement Advisor based on the current configuration and workload. This IOPS analysis also confirms the suitability of hosting the new workload on this storage system

- **Data Rate:** Data rate refers to the speed at which the overall data is transferred to and from the storage system. The analysis considers the historical data rate and considers the observed growth rate to forecast the future data rate. The forecast is plotted to show the data rate with and without the new workload, as illustrated in Figure 4-9.

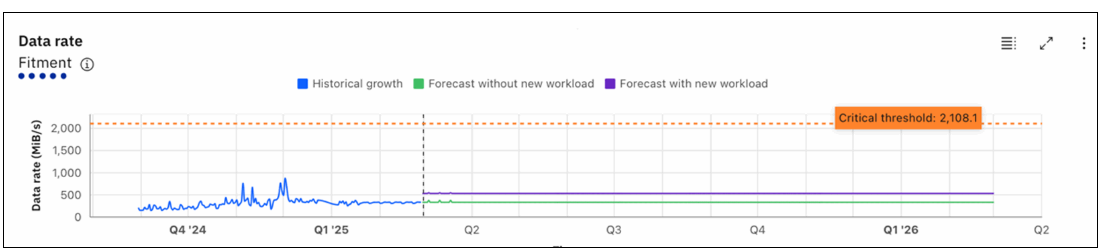


Figure 4-9 Data rate metric analysis

As seen in Figure 4-9, the expected data rate on the target storage are less than the critical threshold of the target storage system. The critical threshold is derived by the Workload placement advisor and is based on the current configuration and the current workload. The data rate analysis also shows that hosting the new workload on this storage system is again suitable.

- **Latency:** Latency is the expected response time for input-output operations, including reads and writes. The analysis compares historical latency on the target storage with the new workload's projected latency. Figure 4-10 shows a comparative latency forecast with and without the new workload.

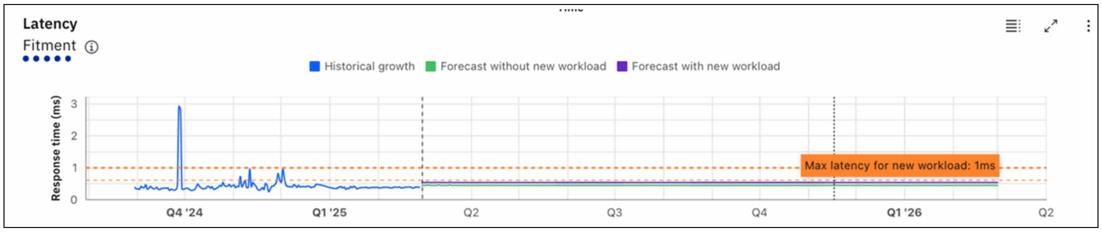


Figure 4-10 Latency analysis

As seen in Figure 4-10, the expected latency with the new workload on the target storage is less than expected and hence the storage system is suitable to host the new workload.

- **CPU Utilization:** This metric represents the usage of all CPU cores available on the target storage system controller. Best practices suggest maintaining CPU utilization below 60%, while IBM Storage Insights considers utilization exceeding 80% a critical alert. The analysis forecasts CPU utilization based on the projected workload, taking into account the existing workload on the target storage system and its anticipated growth. Figure 4-11 illustrates a forecast comparing CPU utilization with and without the new workload.

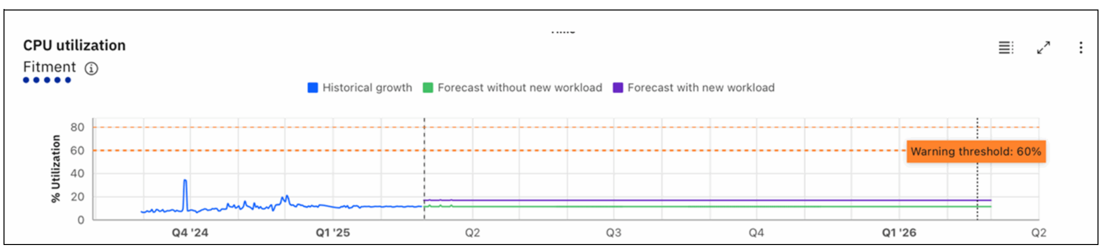


Figure 4-11 CPU utilization analysis

Because the expected CPU utilization on the target storage with the new workload is below 60%, as shown in Figure 4-11, the storage system is deemed suitable for hosting it.

In summary, the Storage Partition Placement Advisor uses artificial intelligence to guide storage administrators in selecting the optimal storage system for new workloads.

4.3 Storage partition migration within a FlashSystem grid

IBM Storage Insights facilitates storage partition migration between storage systems within a FlashSystem grid. As discussed in 4.1, “Storage partition and FlashSystem grid” on page 28, a storage partition is a logical grouping of resources representing a workload. Consequently, planning for storage partition migration often requires administrators to carefully consider storage and system resources to avoid impacting existing workloads. A significant advantage of logically grouping workloads in storage partitions is that the migration technology within the FlashSystem grid supports live workload migration with zero application downtime. This capability introduces complexity in precisely planning the selection of a target system to ensure the migrated workload can be sustained long-term without performance degradation.

This is where the IBM Storage Insights Placement Advisory for storage partition comes into play. As detailed in 4.2, “Workload Placement Advisory for storage partition” on page 29, administrators can leverage historical data and future growth predictions to plan storage partition migration from one storage system to another.

Similarly, when an administrator initiates a migration from Storage Insights, AI-recommended target storage systems are presented with similar explanations (as illustrated in Figure 4-12) to guide the storage migration operation. Administrators need to enable the toggle to view the compatibility score, which provides the AI-based recommendations for the target storage systems.

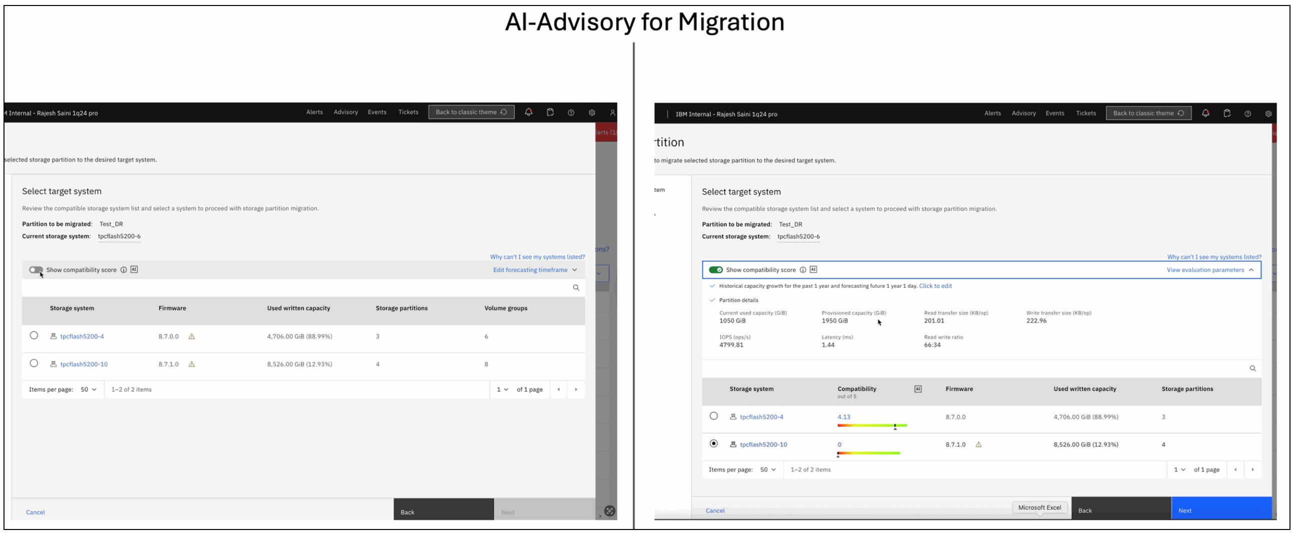


Figure 4-12 Migration Partition with AI advisory

> **Note:** Beginning with FlashSystem version 9.1.0, IBM Storage Insights is expected to offer AI-driven recommendations for target storage system selection during FlashSystem partition migrations initiated from the FlashSystem UI (Forward-looking statements are subject to change).

Beyond user-initiated workload advice, IBM Storage Insights also monitors resources within the FlashSystem grid and proactively provides recommendations to users based on the performance and storage requirements of the systems within that grid. These proactive advisories are visible on the overview dashboard and the FlashSystem grid dashboard, as shown in Figure 4-13 on page 37.

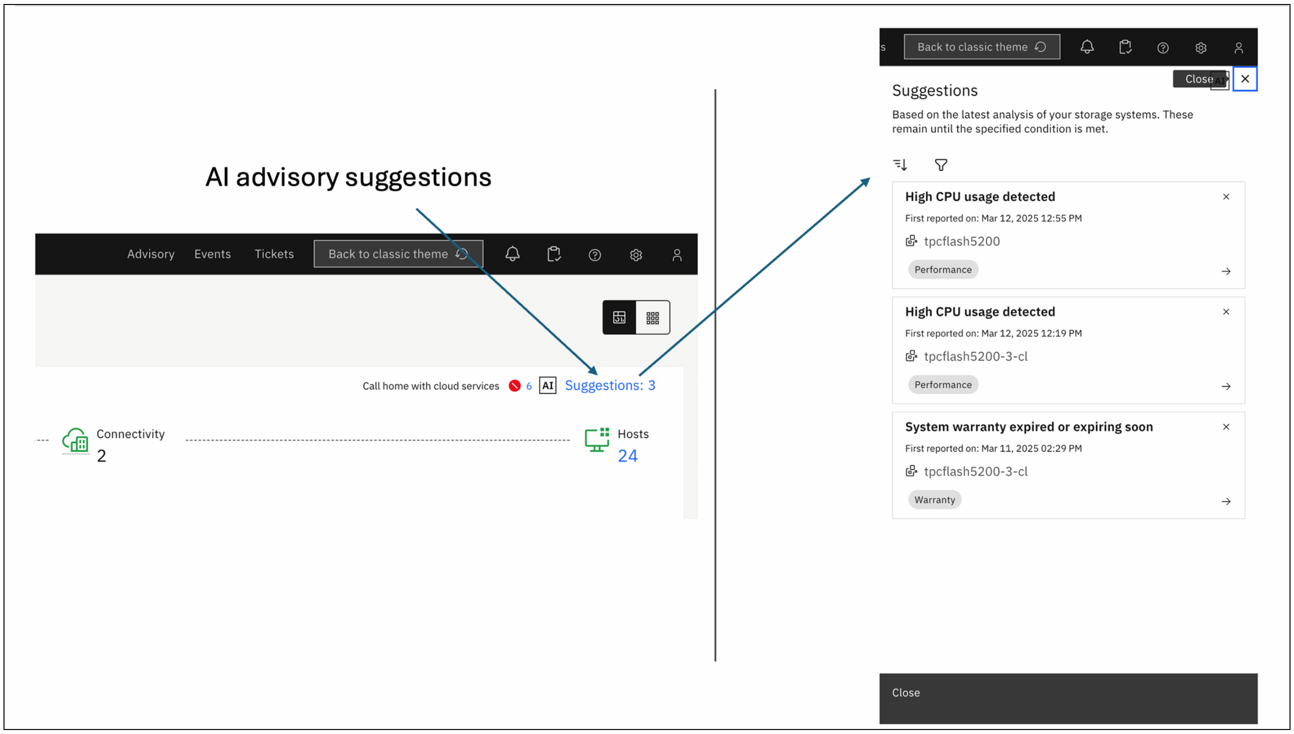


Figure 4-13   AI-driven insights on the overview dashboard

For a deeper understanding of AI-Advisory recommendations, consult the IBM Storage Insights AI-Workload Advisory documentation.

5

System health advisory for FlashSystem grid

IBM Storage Insights advisories are designed to explain problems and offer informed recommendations for solutions. Currently, IBM Storage Insights supports health check advisories. Similarly, artificial intelligence powers the creation of *AI suggestions*. These provide customers with recommendations, such as storage partition migration, based on the storage system's capacity, maintenance status, and performance conditions. These AI suggestions are specifically for block storage systems within a FlashSystem grid.

In this chapter we discuss System health advisory for FlashSystem grids. This chapter has the following sections:

5.1 Conditions of the AI suggestions

AI suggestions are generated after the daily inventory data collection process is complete. Currently, IBM Storage Insights supports AI suggestions for three conditions:

- **Warranty and Maintenance:** An AI suggestion indicating that the system warranty is expired or expiring soon will be generated for a block storage system if it is part of an IBM Storage FlashSystem grid and its warranty and maintenance are set to expire within the next 30 days.
- **Capacity:** An AI suggestion for insufficient storage capacity to sustain growth will be generated for a block storage system if it is part of a FlashSystem grid and its forecasted used capacity is projected to reach or exceed 80% of its total capacity within the next 30 days.
- **Performance:** An AI suggestion will alert users to an expired or soon-to-expire system warranty for a block storage system within an IBM Storage FlashSystem grid if its warranty and maintenance are set to expire within the next 30 days.

For block storage systems experiencing a ransomware threat, AI suggestions will not be generated, and any existing AI suggestions will be removed. Ransomware threat is a critical alert which needs to resolved first, as it may impact the criteria data points used for AI suggestions.

Upon completion of each probe, the AI suggestions will be revalidated, and any suggestion whose conditions are no longer met will be deleted.

5.2 Accessing the AI suggestion

The AI suggestion will be accessible from three places.

- The Overview page (L2 page) of block storage systems (Figure 5-1). The AI suggestion count represents the total number of current AI suggestions generated for block storage systems within this tenant (IBM Storage Insights instance).

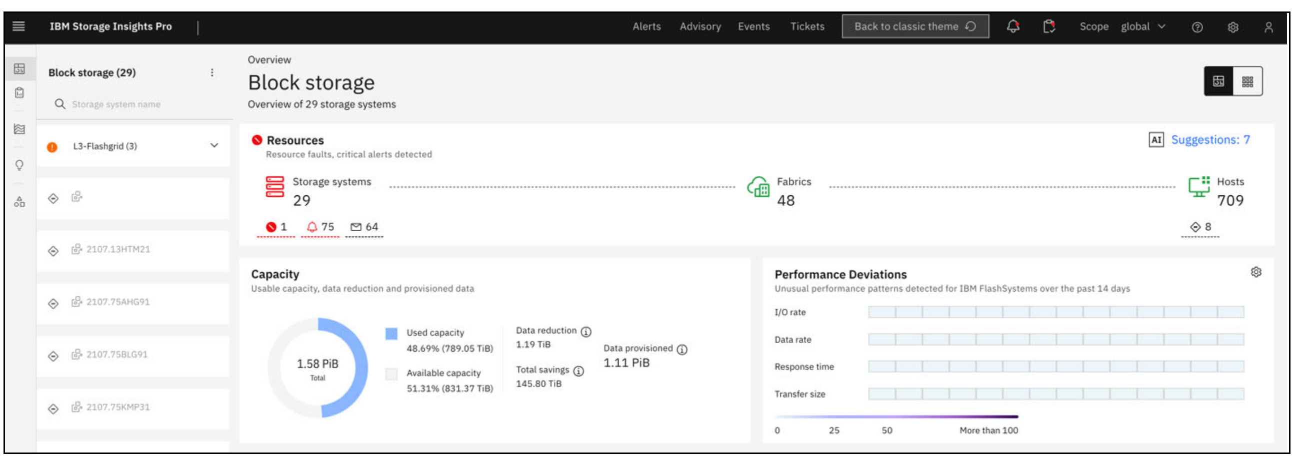


Figure 5-1 Block storage overview AI suggestion

- The Overview page of the FlashSystem grid (Figure 5-2). The AI suggestion count will be the count of all current AI suggestions generated for block storage system generated under this FlashSystem grid.

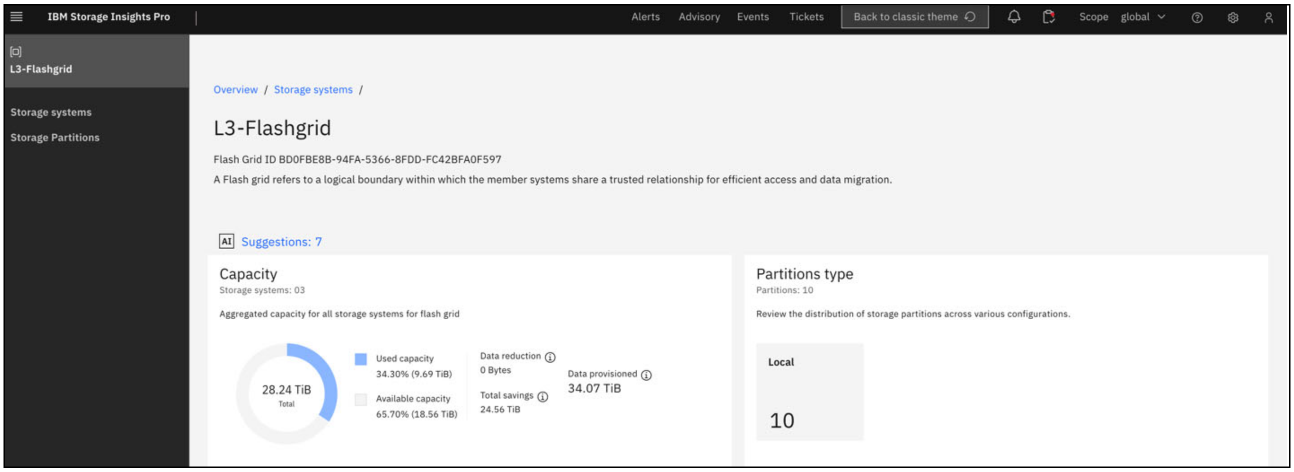


Figure 5-2 FlashSystem grid overview the AI suggestion

- The Details page of the block storage system (Figure 5-3). The AI suggestion count represents the total number of current AI suggestions generated for this block storage system.

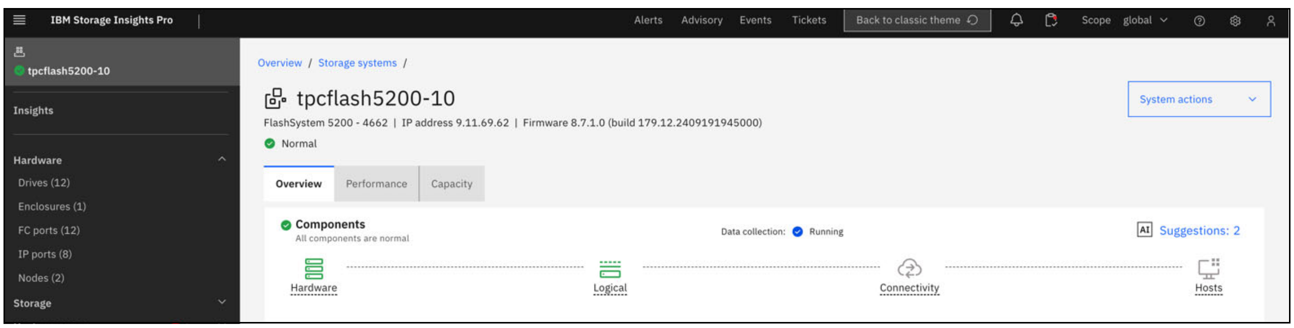


Figure 5-3 IBM FlashSystem storage overview AI suggestion

5.2.1 Summary view of the AI suggestion

Clicking the **AI suggestion count** from any location will open the right panel, displaying a summary of the AI suggestions. See Figure 5-4.

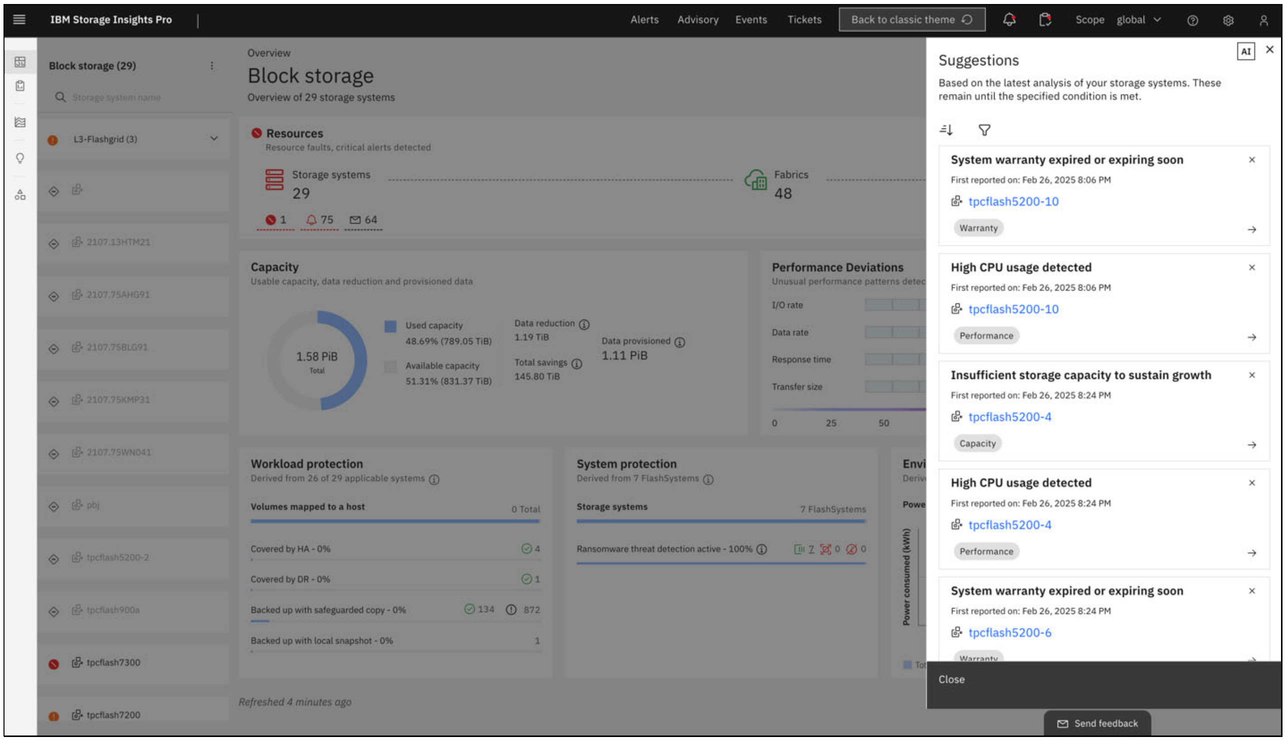


Figure 5-4 Summary view of AI suggestion

As shown in Figure 5-5 on page 42, the AI suggestion list summary view can be accessed via an individual FlashSystem details page.

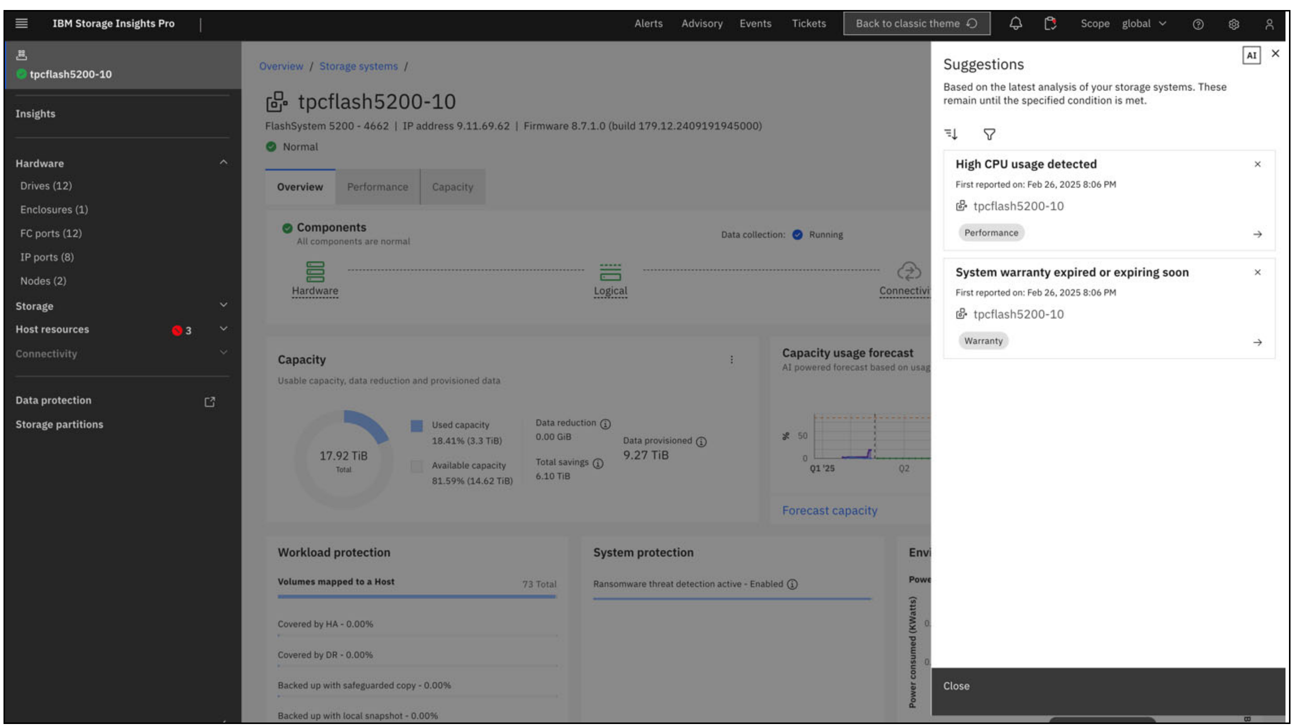


Figure 5-5 Summary view of AI suggestion for individual IBM flash system

The summary list view provides the capability to sort AI suggestions by creation date as well as by category.

5.2.2 Details view of AI suggestions

When you open an AI suggestion, you will see its detailed view. The **Overview** tab analyzes the AI suggestion and provides a link for deeper understanding of the condition. The **Recommendation** tab lists recommendations relevant to this AI suggestion and links to available actions.

- Maintenance AI suggestions: Figure 5-6 on page 43 is an example of AI suggestion for when storage system maintenance expired or expiring in next 30 days.

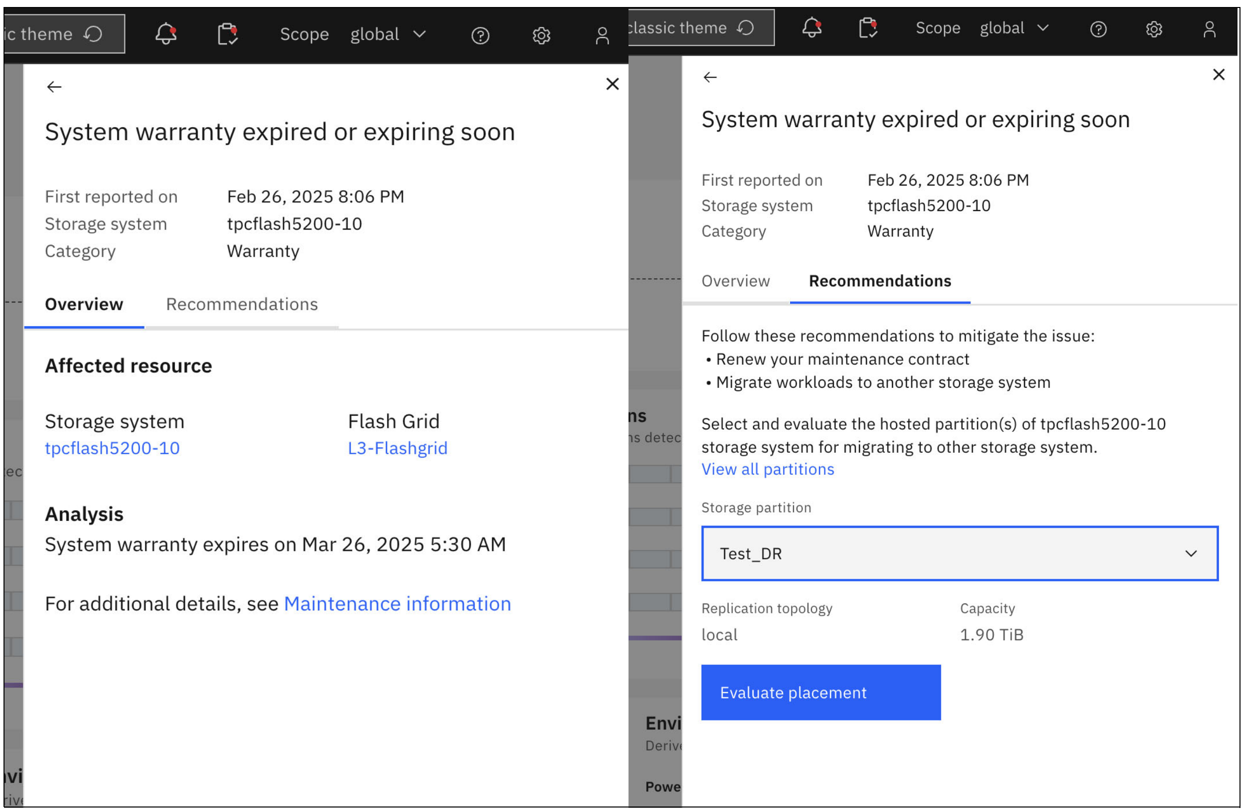


Figure 5-6 Detail view of a maintenance AI suggestion

Recommendations offer various ways to address the issue. The user can renew the maintenance contract to continue using the current storage system with valid support. Alternatively, the user can migrate workloads from the current storage system to another system with a valid maintenance contract to ensure the security of the workloads. To assist with this, the user can utilize the “Evaluate Placement” feature in IBM Storage Insights.

- Capacity AI suggestions: Figure 5-7 on page 44 is an example of AI suggestion for when capacity forecasting reaching 80% in next 30 days.

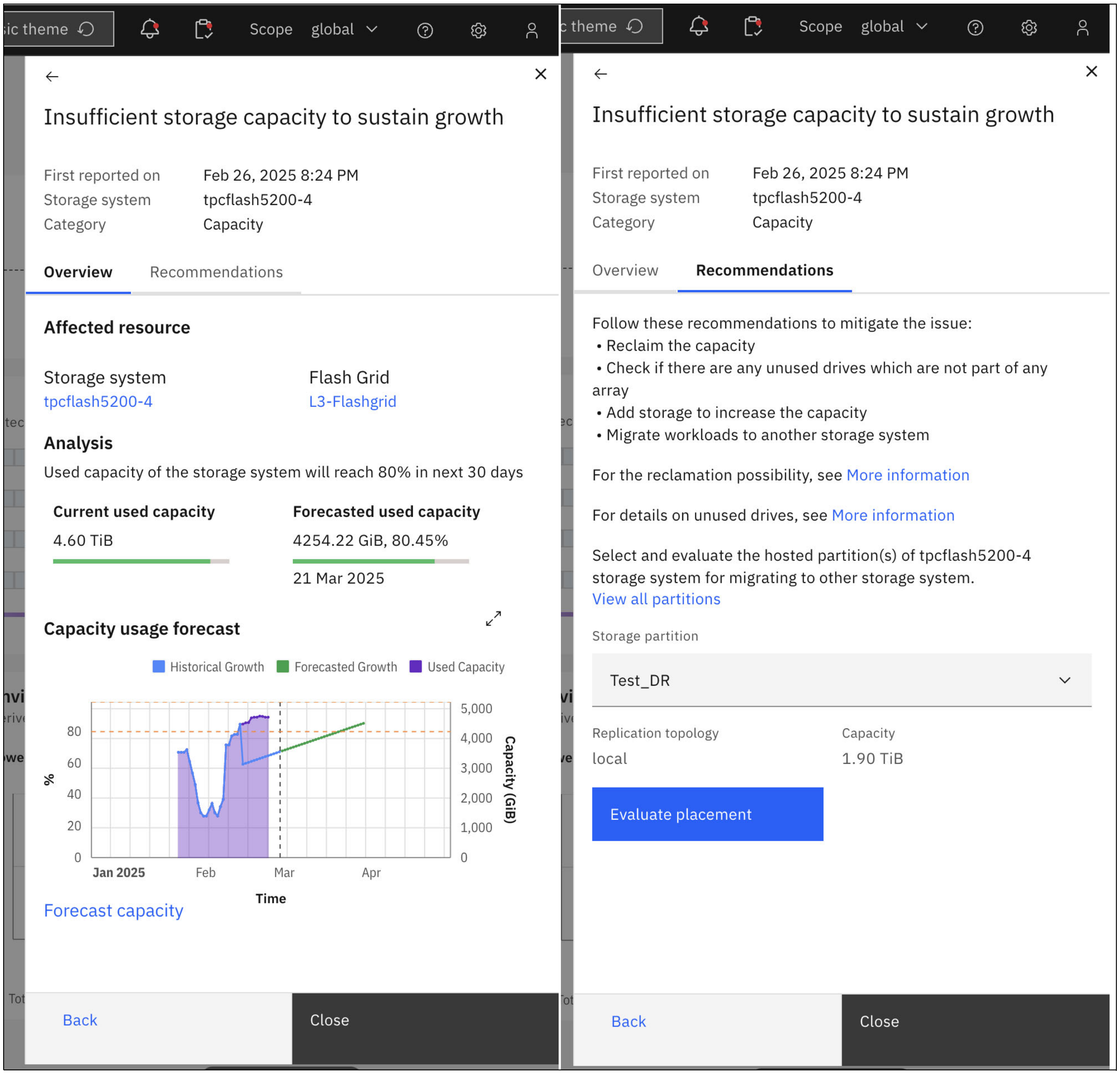


Figure 5-7 Detail view of a capacity AI suggestion

Recommendations offer various approaches to resolve the issue. The user can utilize the Reclamation feature to identify reclaimable volumes and storage. To increase overall capacity, the user can add additional storage. Alternatively, the user can migrate workloads from the current storage system to another system with greater storage capacity to ensure workload security. For this purpose, the “Evaluate Placement” feature in IBM Storage Insights can be used.

- Performance AI suggestions: Figure 5-8 on page 45 is an example of AI suggestion for when CPU usage reached 75% for the 95th percentile of last two days sample data.

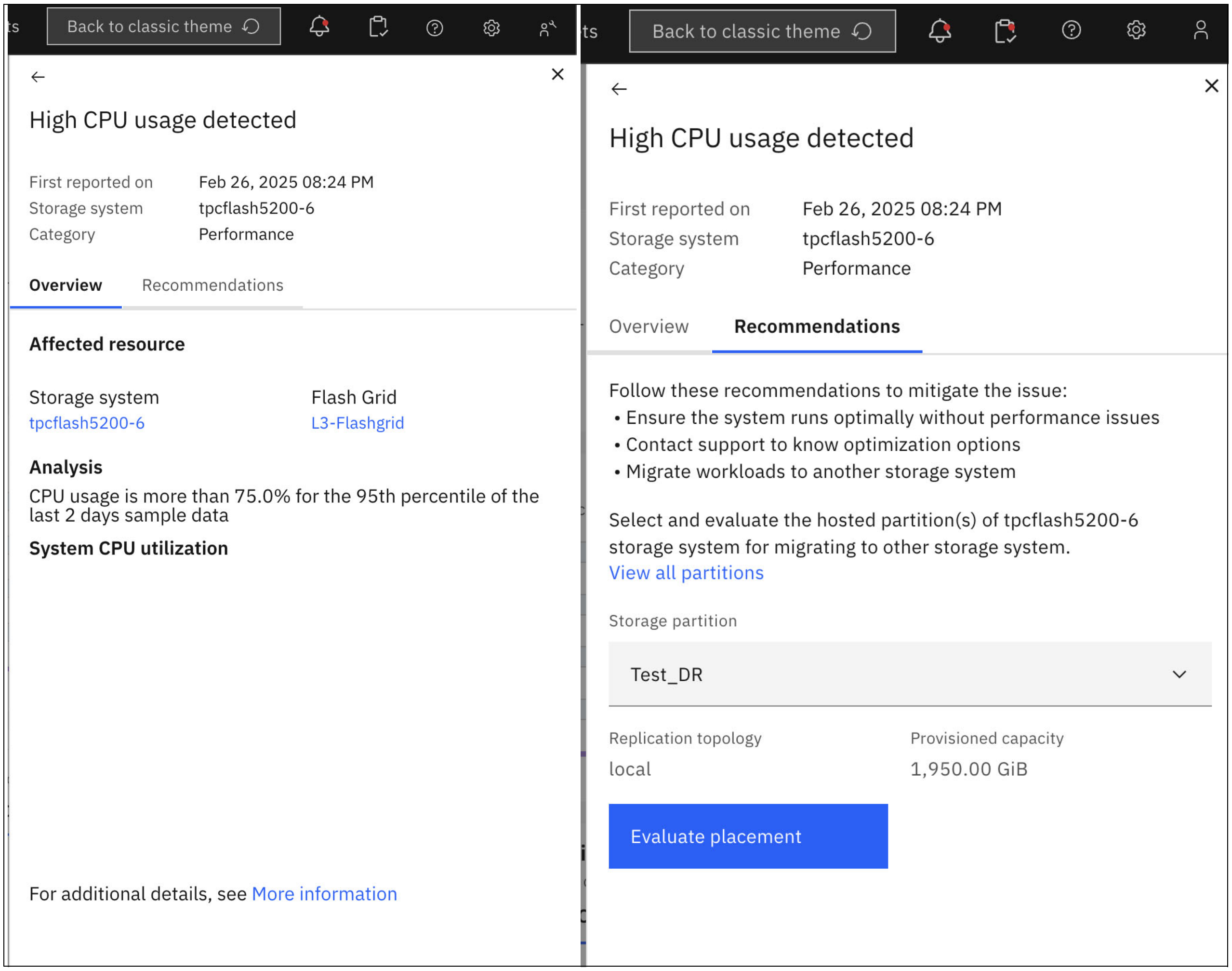


Figure 5-8 Detail view of a performance AI suggestion

Recommendations offer various ways to resolve the issue. Users need to implement appropriate solutions to ensure the system operates without performance problems. Alternatively, users can migrate workloads from the current storage system to another system with greater processing power to safeguard the workloads. To assist with this, the “Evaluate Placement” feature in IBM Storage Insights can be utilized.

6

Workload Anomaly Detection and Ransomware Threat Detection

Leveraging IBM Storage Insights and IBM Storage FlashSystem products provide Workload Anomaly Detection and Ransomware Threat Detection. These IBM-developed functionalities are specifically designed to aid system administrators in proactively protecting their systems against workload anomalies and ransomware attacks.

This chapter covers Workload Anomaly Detection (WAD) and Ransomware Threat Detection (RTD) and has the following sections:

6.1 Introduction

Workload Anomaly Detection (WAD) and Ransomware Threat Detection (RTD) are built-in functionalities of IBM Storage FlashSystem products, integrated with IBM Storage Insights. They empower system administrators to safeguard their systems against workload anomalies and ransomware attacks.

While Ransomware Threat Detection provides finer and almost real-time detection, it poses specific prerequisites regarding the system configuration over Workload Anomaly Detection, as explained later in this chapter. Table 6-1 summarizes the support conditions Workload Anomaly Detection and Ransomware Threat Detection concerning IBM Storage Virtualize versions and FlashCore Module Gen 4 (FCM4):

Table 6-1 WAD and RTD support conditions concerning IBM Storage Virtualize versions and FCM4

IBM Storage Virtualize version	FlashSystem with at least one FCM4 drive	FlashSystem with no FCM4 drive
Below 8.6.0.0	No WAD and RTD are supported.	No WAD and RTD are supported.
8.6.0.0 to below 8.6.3.0	Only WAD is supported.	Only WAD is supported.
8.6.3.0 or later	Only RTD is supported.	Only WAD is supported.

This chapter briefly explains those functionalities' key concepts. You can refer to the official product publication listed below for more details:

- Alerts for workload anomaly detection
- Alerts for ransomware threat detection

> **Note:** Both Workload Anomaly Detection and Ransomware Threat Detection require that IBM Storage Insights (Pro) be configured to monitor storage systems. While the storage systems can be connected using Cloud Call Home (CCH) or via data collectors for Workload Anomaly Detection, only CCH is supported for Ransomware Threat Detection.

6.2 Workload Anomaly Detection

Workload Anomaly Detection was initially introduced with IBM Storage Virtualize 8.6.0 and has since evolved. This feature calculates *Shannon entropy* (also known as entropy) and compressibility on incoming write I/O requests at both the canister node and volume levels. However, due to its computationally intensive nature during the cache destage phase, it samples one write request per hundred I/Os to mitigate performance impacts.

FlashSystem metrics, including Shannon entropy and compressibility, are collected from the storage system and sent to IBM Storage Insights. IBM Storage Insights then runs an algorithm to identify workload anomalies and generate an alert, as depicted in Figure 6-1 on page 49.

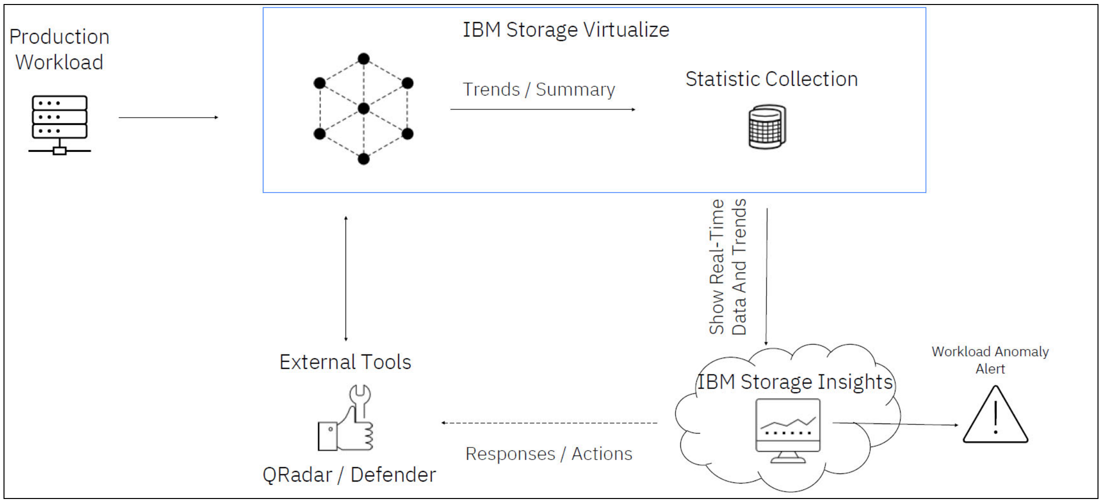


Figure 6-1 Workload Anomaly Detection architectural overview

Whenever a workload anomaly alert is triggered, an email notification will be sent to the configured email addresses. This email will contain details about the anomalies detected in your storage environment and direct links to the corresponding alert within IBM Storage Insights.

6.2.1 Workload Anomaly Detection with FlashCore Module Gen 1, 2, and 3

On FlashSystem systems equipped with FlashCore Module Gen 1, 2, and 3 drives, the Workload Anomaly Detection inference engine uses those entropies and volume-level statistics to shorten the detection of possible workload anomalies to less than an hour.

6.3 Ransomware Threat Detection

Ransomware Threat Detection has evolved from Workload Anomaly Detection in IBM Storage Virtualize 8.6.3, and it is only applicable to the systems that satisfy the following prerequisites:

- An IBM Storage FlashSystem system equipped with FlashCore Module Gen 4 (or later) drives.
- The system is configured with standard pools containing a single Dynamic RAID (DRAID) array and configured on IBM Storage Virtualize 8.6.3 and later.
  - When Data Reduction Pools (DRPs) are configured, only Fully Allocated volumes can be detected by Ransomware Threat Detection; other volume types, such as Deduplicated volumes, are not detected by Ransomware Threat Detection.
- The DRAID array is configured with FCM4 only, whose drive firmware level is 4.1 and later.
- The system is connected to IBM support via Cloud Call Home (CCH) and monitored by an IBM Storage Insights Pro instance.
  - If the system is connected to an IBM Storage Insights Pro instance via data collectors, it does not satisfy the prerequisite condition.
- A minimum of 128 GB memory per node.
  - The CCH integration of IBM Storage Insights imposes this requirement.

> **Note:** IBM Storage FlashSystem 5000 series and 5200 products do not satisfy these requirements because they cannot be equipped with FCM4 drives. IBM Storage FlashSystem 5300 can meet the requirement only when configured with a minimum of 128 GB memory per canister node.

The “Ransomware Threat Detection” column in Figure 6-2 designates the Ransomware Threat Detection-enabled systems on IBM Storage Insights classic UI.

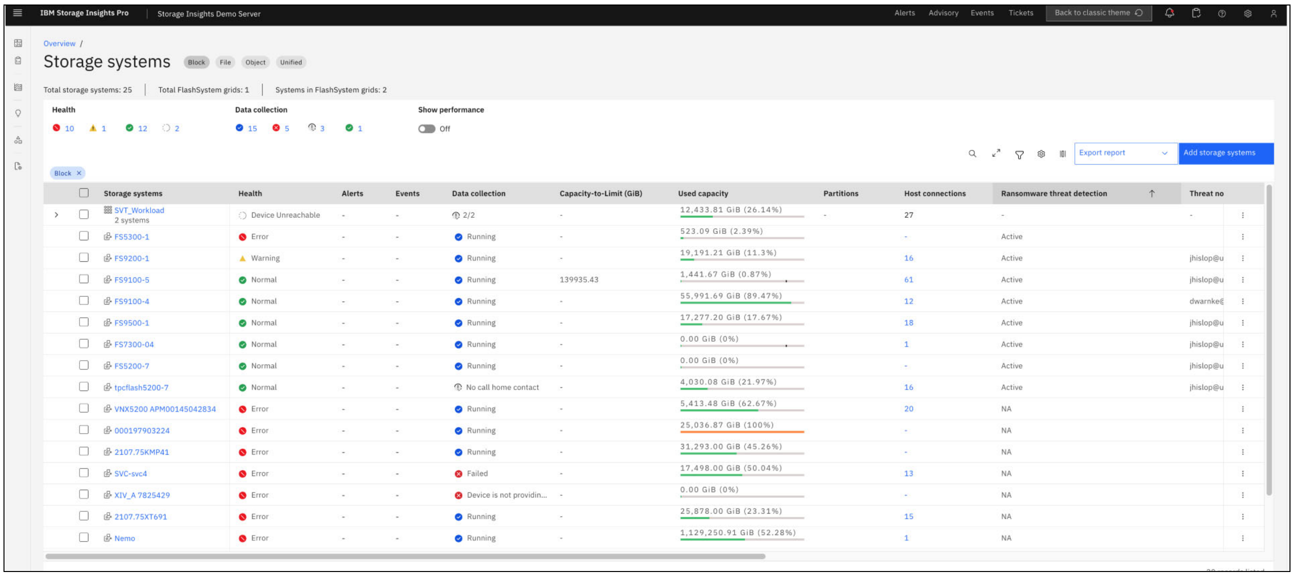


Figure 6-2 Ransomware Threat Detection-enabled systems on IBM Storage Insights

By utilizing the FCM4’s advanced hardware functionality, Ransomware Threat Detection drastically shortens the time needed to detect possible ransomware attacks to less than a minute, nearly real-time.

You can enable and disable alerts for RAD by following the procedure explained in Enabling and disabling alerts for ransomware threat detection.

You will receive an email notification at the configured email addresses whenever a ransomware threat alert is triggered. This email includes details about the potential ransomware attack on storage resources such as volumes and volume groups and provides a link to the corresponding alert in IBM Storage Insights. Ransomware alerts are configured by default for eligible devices.

6.3.1 Ransomware Threat Detection with FlashCore Module Gen 4 or later

The FCM4 is designed to provide even greater support for the AI Inference Engine than earlier generations. Equipped with two real-time processing units, it enables the Summarizer to operate independently and rapidly calculate essential statistics, complementing those processed by Workload Anomaly Detection, as illustrated in Figure 6-3 on page 51.

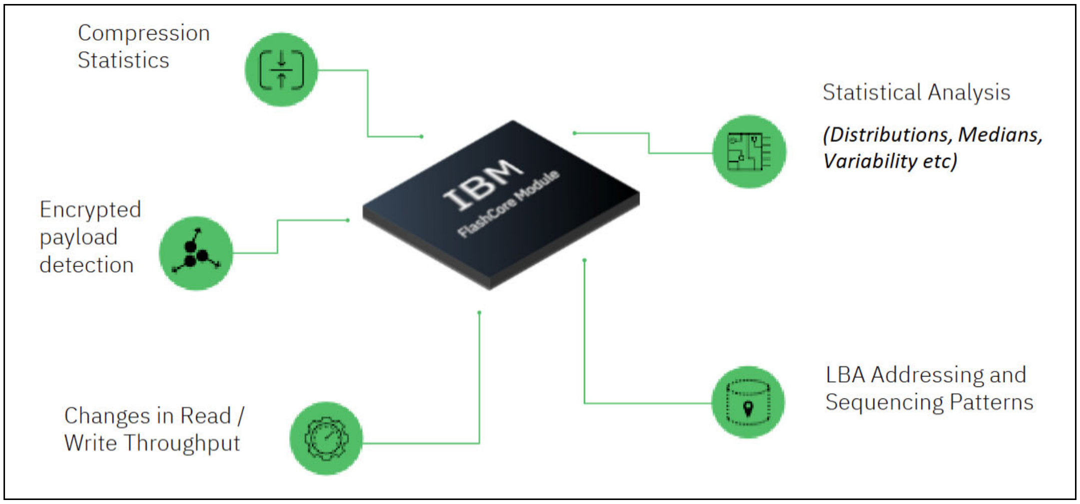


Figure 6-3 Statistics calculated by FCM4

As depicted in Figure 6-4, the Aggregator collects all the information from the FCM4 drives every two seconds and processes the gathered statistics on a per-volume basis. The information is then passed to the Revealer, an Inference Engine running in the IBM Storage Virtualize. it examines all the statistics and decides whether an anomaly has occurred or if ransomware appears in nearly real-time.

Although a volume’s Logical Block Address (LBA) is mapped onto multiple FCM4 drives comprising a DRAID array, the algorithm can aggregate and adjust the dispersed LBA addressing and sequencing patterns from member drives.

IBM calls this FCM4’s design concept *computational storage*, which enables Ransomware Threat Detection. See Figure 6-4.

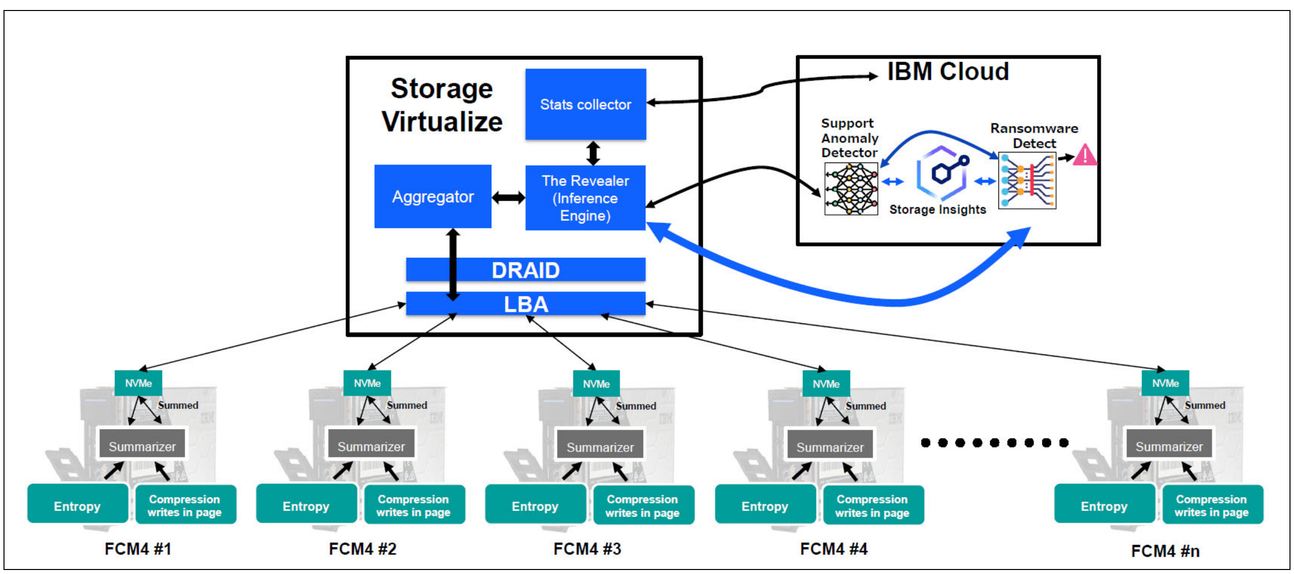


Figure 6-4 FlashSystem Ransomware Threat Detection with FCM4 conceptual model

The Ransomware Threat Detection inference engine’s model is being trained remotely using machine learning (ML) to lower the probability of a detection result being found as a *false positive*, as explained in Figure 6-4. The model is extensively trained with database OLTP-type workloads.

Once IBM decides to release it, a new ML-trained algorithm can be delivered via patching mechanisms to systems running IBM Storage Virtualize 8.7.0 and later. Then, those systems can be updated dynamically without restarting canister nodes.

6.3.2 File system awareness

IBM keeps enhancing Workload Anomaly Detection and Ransomware Threat Detection to make those functionalities more reliable. An example of those enhancement efforts is *File System Awareness*, which was introduced in IBM Storage Virtualize 8.7.0.

All the volumes on FlashSystem running IBM Storage Virtualize 8.7.0 and later are automatically scanned twice a day to detect the file system type. Once detected, the new file_system attribute for a volume is filled with the corresponding value, as shown in the following **lsvdiskanalysis** command output (Example 6-1 on page 52).

Example 6-1 svdiskanalysis command output

```
fs9500:~ # lsvdiskanalysis 0
id 0
name vdisk0
state sparse
started_time 240425173030
analysis_time 240425173030
capacity 100.00GB
thin_size 0.00MB
thin_savings 0.00MB
thin_savings_ratio 0
compressed_size 0.00MB
compression_savings 0.00MB
compression_savings_ratio 0
total_savings 0.00MB
total_savings_ratio 0
margin_of_error 0
file_system ext4,xfs
```

The inference engine uses this file system type to improve the detection probability, such as placing multiple virtual machine images (VMDKs) on a single VMFS datastore, a volume mapped from FlashSystem to VMware ESXi hosts.

Note: Data read on the volume is only used to determine file system type and is *not* sent to IBM Storage Insights for processing.

6.4 Alerts on IBM Storage Insights

Once detected, you can access alerts as explained in the following two sections.

6.4.1 Accessing workload anomaly threat alerts

You can access alerts for Workload Anomaly Detection on IBM Storage Insights as follows:

- Modern UI: Bell icon on the overview page. Locate Workload anomaly detected alert tile.
- Classic UI: You can access workload anomaly alerts through the following locations:

- Dashboards → Alerts.
- Resources → Block storage systems → Alerts.
- Alerts tab in the General section of the storage system's details page.
- Volumes tab in the Internal Resources section of the storage system's details page.

Alert details can be confirmed as follows:

- Modern UI: On the overview page, click the bell icon and locate the **Workload anomaly detected** alert tile page to view similar alert details.
- Classic UII: Double-click the alert in **Dashboards** → **Alerts** to view related storage systems, affected volume(s) table, performance charts (Read I/O Rate, Write I/O Rate, Total I/O Rate), and recommended mitigation actions.

6.4.2 Accessing ransomware threat alerts

You can access alerts for Ransomware Threat Detection on IBM Storage Insights Pro as follows:

- Modern UI: You can access ransomware alerts for either volume or volume groups through the following locations:
  - Click **View threat alerts** from the ransomware threat notification.
  - Bell icon on the overview page. Locate the ransomware threat alert tile.
- Classic UI: You can access ransomware alerts for volume groups through the following locations:
  - Click **Go to Alerts** from the ransomware threat notification.
  - **Dashboards** → **Alerts**.
  - **Resources** → **Block storage systems** → **Alerts**.
  - **Alerts** tab in the **General** section of the storage system's details page.
  - **Volume** and **Volume Group** tab in the **Internal Resources** section of the storage system's details page.

Alert details can be confirmed as follows:

- Modern UI: On the overview page, click the bell icon and locate the ransomware alert tile page to view similar details.
- Classic UI: Double-click the alert in **Dashboards** → **Alerts** to view related storage systems, volume group tables, performance charts (Read I/O Rate, Write I/O Rate, Total I/O Rate), and recommended mitigation actions.

Within IBM Storage Insights, Ransomware Threat Detection encompasses volume groups in addition to individual volume-based alerts. When an event is detected at the group level, IBM Storage FlashSystem sends an alert to IBM Storage Insights, allowing for quick identification and management of compromised volume groups rather than addressing each volume individually.

> **Note:** Real-time diagnostics for volume groups is supported on FlashSystem, which is equipped with FCM4 drives and runs IBM Storage Virtualize 8.7.2.0 or later.

As depicted in Figure 6-5 on page 54, IBM Storage Insights can mark snapshots created after detection as *Compromised*.

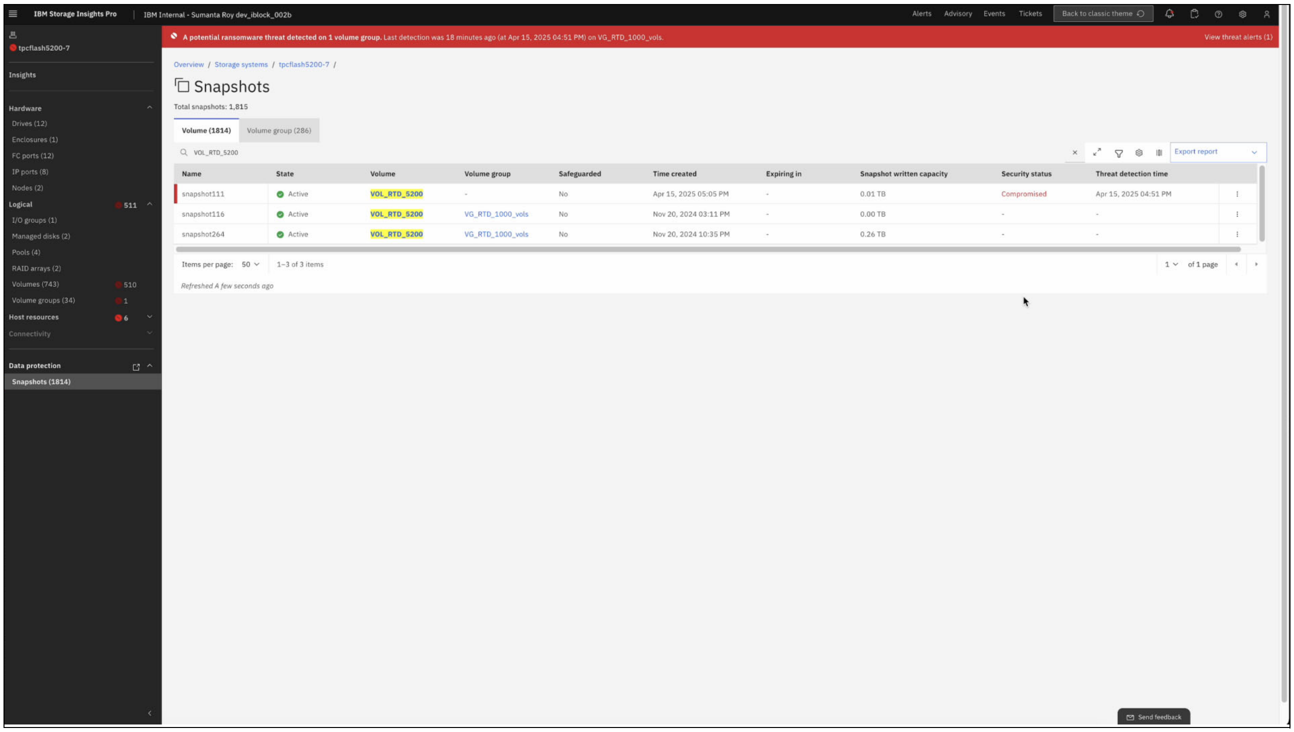


Figure 6-5 Volume snapshots marked compromised on IBM Storage Insights

6.4.3 How to react to alerts

Figure 6-6 on page 55 shows alert details for a potential ransomware detection.

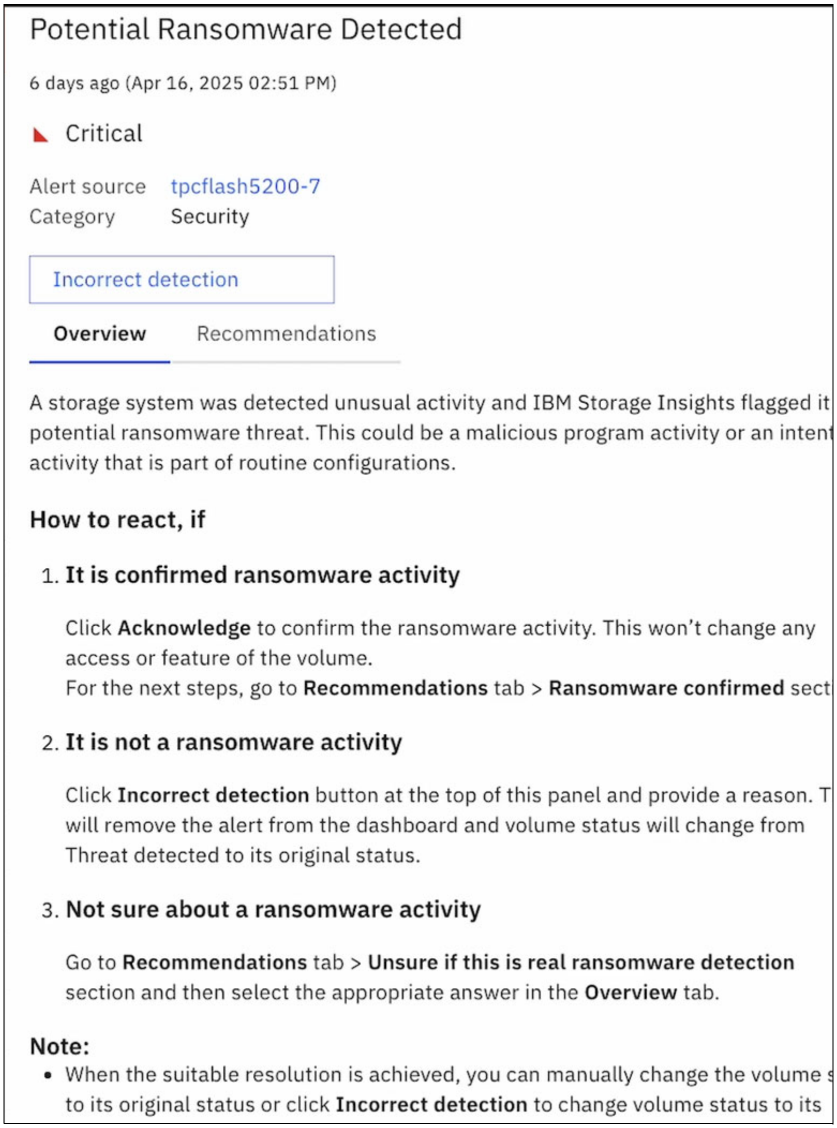


Figure 6-6 Alert example of a potential ransomware detection on IBM Storage Insights

- Click on **Acknowledge** if this is confirmed ransomware activity, then go to the **Ransomware confirmed** section under **Recommendation** on this panel. Once a volume has been mitigated or fixed following a ransomware threat, it (or its volume group) should be marked as Acknowledged. To do this, right-click the desired volume or volume group showing the "Ransomware threat detected" status, and then click **Mark Status as Acknowledged**.
- Click on **Incorrect detection** if this is confirmed not a ransomware activity and provide a reason. This will remove the alert from the dashboard, and the volume status will change from Threat detected to its original status. This informs IBM that this detection is a false positive to improve the model by further training.
- If you are not sure about ransomware activity on this alert, follow the following instructions shown under the **Recommendations** on this subsequent panel **Unsure if this is real ransomware detection**:
  - Contact your operating system or hypervisor team to verify whether any configuration changes were made when the alert was triggered. The following configuration changes (but are not limited to) can trigger the alert:

- Enabling application compression.
- Enabling application encryption.
- A significant increase in the data stored on the volume. For example, migrating data from other volumes, especially if the new data is less compressible than the data before migration activity.
- Restoring from backups or replicated environments.
- Installing new applications and workloads.

– If none of the above activities occurred, go to the **Ransomware confirmed** section, as shown in Figure 6-7, and follow the instructions provided there.

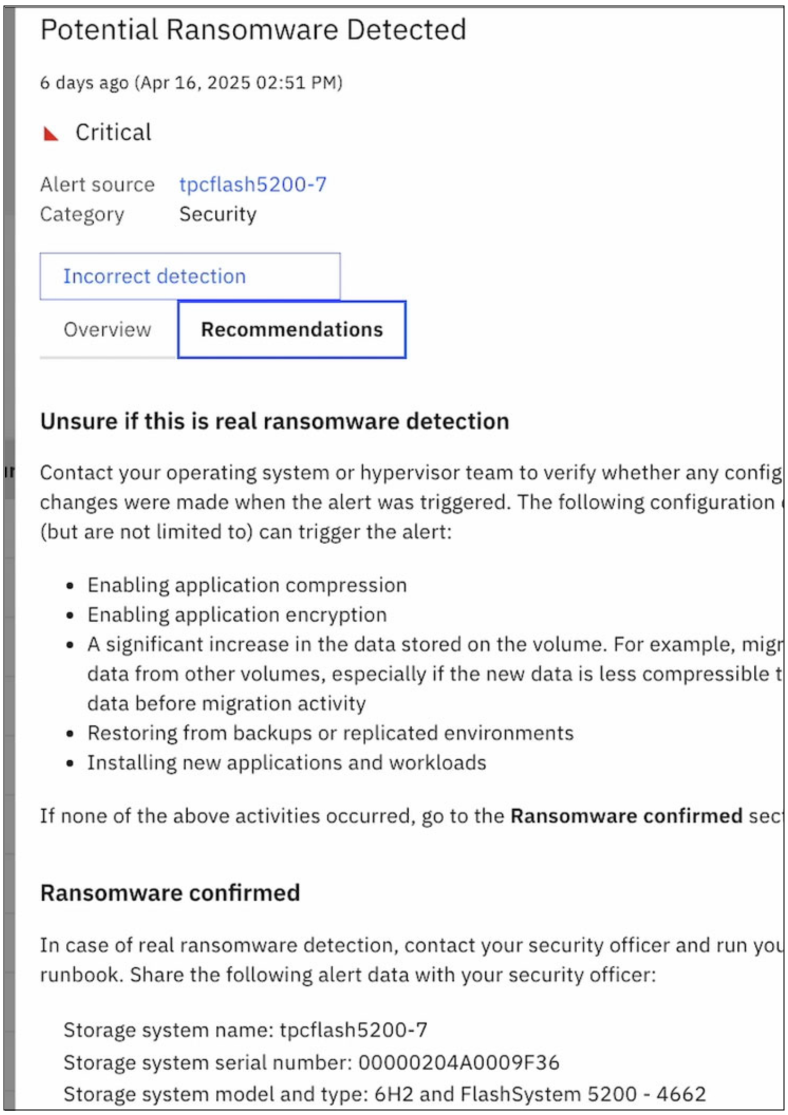


Figure 6-7 Ransomware confirmed section for an alert example

7

Observability Chatbot for IBM Storage Insights

In large-scale enterprise environments, efficient monitoring and visibility into a storage infrastructure are critical for ensuring system reliability, performance, and operational excellence. Although platforms like IBM Storage Insights provide powerful observability and AIOps (Artificial Intelligence for IT Operations) capabilities, many of their advanced features are accessed through REST APIs or require manual dashboard navigation, which introduces friction to day-to-day operations.

To streamline this process and enhance observability, the IBM Storage Insights team developed and open-sourced an AI-powered Observability Chatbot. This chapter introduces the key features, high-level architecture, deployment, and technical implementation of the chatbot, which serves as a conversational interface to IBM Storage Insights.

This chapter has the following sections:

7.1 Introduction

With the increasing complexity of hybrid and multi-cloud storage environments, administrators require streamlined tools for visibility and control. Traditional interfaces often present steep learning curves or require manual effort to retrieve insights. This section introduces the chatbot's purpose, the challenges it addresses, and its key functions.

7.1.1 What is the Observability Chatbot?

Observability plays a crucial role in modern IT operations, offering visibility into infrastructure, applications, and service health. IBM Storage Insights is a comprehensive observability and AIOps platform that enables storage administrators and Site Reliability Engineers (SREs) to efficiently monitor, analyze, and manage enterprise storage systems. However, interacting with monitoring platforms often requires manual API requests or complex dashboard navigation, which can be time-consuming.

To address these challenges, IBM Storage Insights includes the Observability Chatbot, which is an AI-driven conversational assistant that integrates with IBM Storage Insights APIs. This chatbot, which is built with IBM watsonx and large language models (LLMs) like Granite and Llama, enables users to query IBM Storage Insights by using natural language in English, eliminating the need for complex manual interactions.

Through advanced prompt engineering and API integration, the chatbot intelligently detects user intent, extracts relevant entities, and dynamically interacts with the IBM Storage Insights API to provide real-time storage system alerts, metrics, and notifications. The chatbot acts as a natural language interface for IBM Storage Insights, improving observability, reducing operational complexity, and enhancing troubleshooting efficiency.

7.1.2 Key features

The chatbot enhances observability by providing the following functions:

- Natural language understanding: Converts user queries into structured API requests to ensure seamless interaction with IBM Storage Insights APIs.
- Entity recognition: Extracts storage-specific metadata (for example, storage system names, alert types, and capacity metrics).
- REST API integration: Fetches storage alerts, notifications, system status, and more through secure authentication and structured API calls.
- Conversational memory: Maintains context across multiple interactions for better response accuracy.

The chatbot interface is designed to provide an intuitive, conversational experience for interacting with IBM Storage Insights APIs. The UI consists of multiple sections that guide the user through authentication, configuration, conversation initiation, and history retrieval.

7.1.3 UI experience

This section describes the various views of the UI, which are as follows:

- Login window: In this window, users submit their User ID, Tenant ID, and API Key to register, as shown in Figure 7-1.

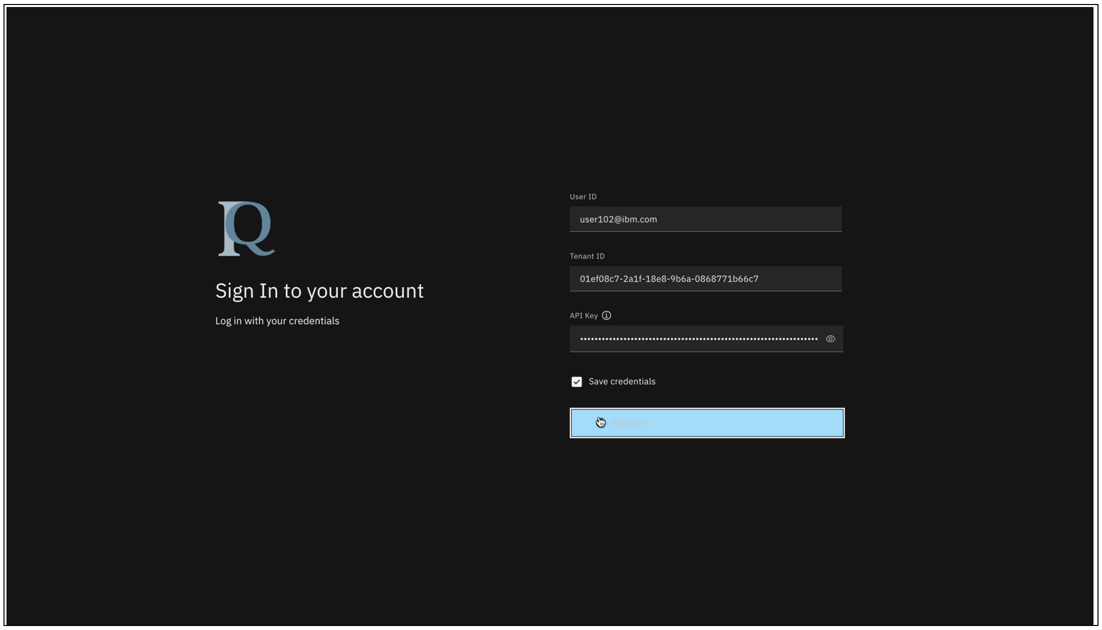


Figure 7-1 Login window

  If you have not generated an API key, see Manage REST APIs.

- The chat dashboard, which is shown in Figure 7-2.

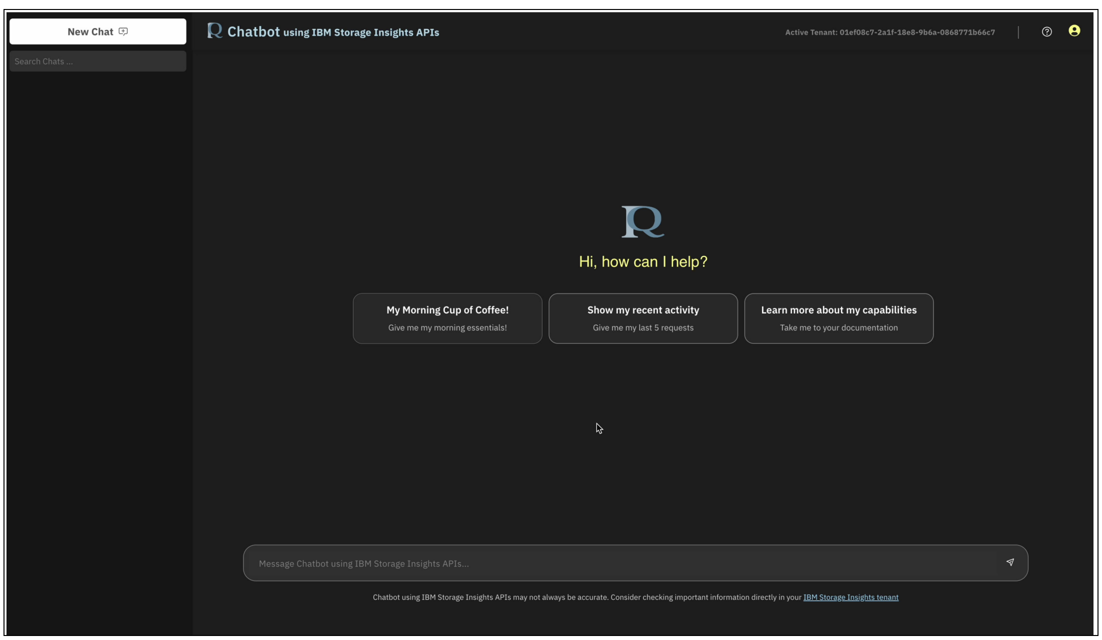


Figure 7-2 Chat dashboard

The chatbot interface has the following features:

- Chat window: Type your queries into the input bar at the bottom of the chat window and press Enter. The chatbot's response appears in the main conversation area.
- Chat History pane: Displays previous chat sessions, as shown in Figure 7-3. You can revisit previous chat threads for context continuity.

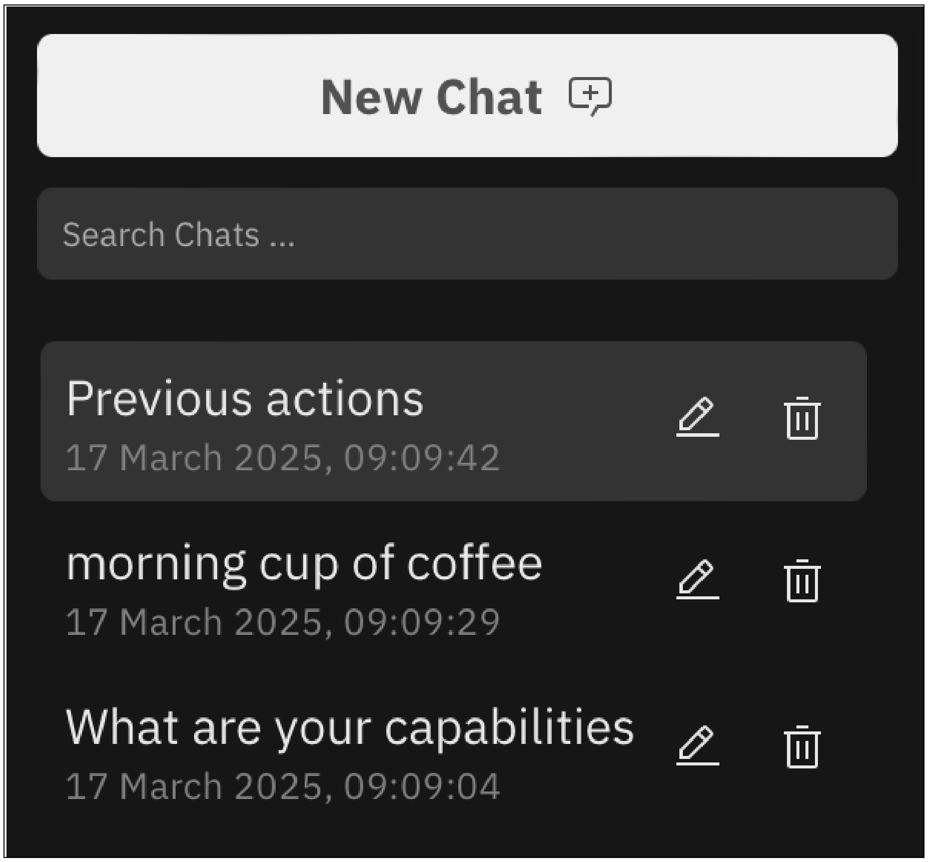


Figure 7-3 Chat history

- Routine Cards: When a new chat session begins, the chatbot presents three routine options, as shown in Figure 7-4:
  - Morning Cup of Coffee: Shows your morning essentials, which is a highlight of all critical alerts and notifications on your storage systems.
  - Recent Activity: Lists the last five previous actions.
  - Capabilities: Highlights chatbot features.

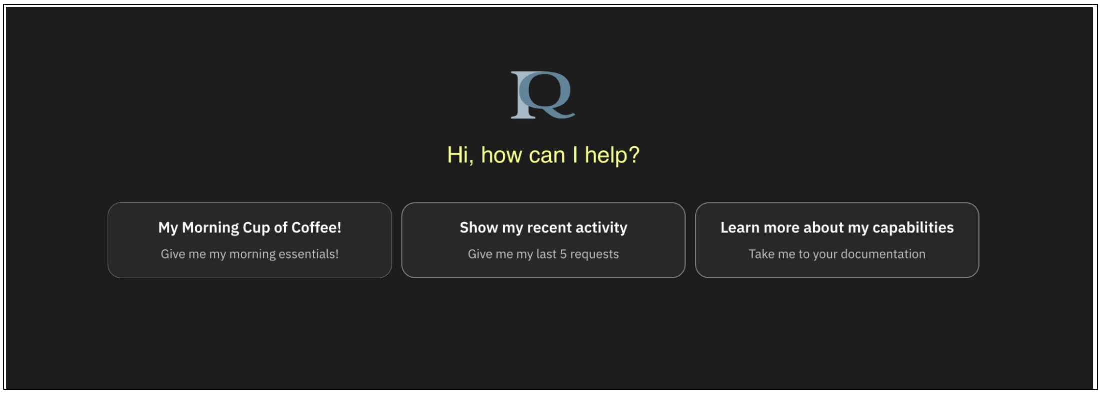


Figure 7-4 Routine cards

To generate the view that is shown in Figure 7-5, click **Show my recent activity** from the routine cards that appear at the start of a new chat session. This action triggers the Previous Actions routine, which shows a curated list of the user's recent API interactions.

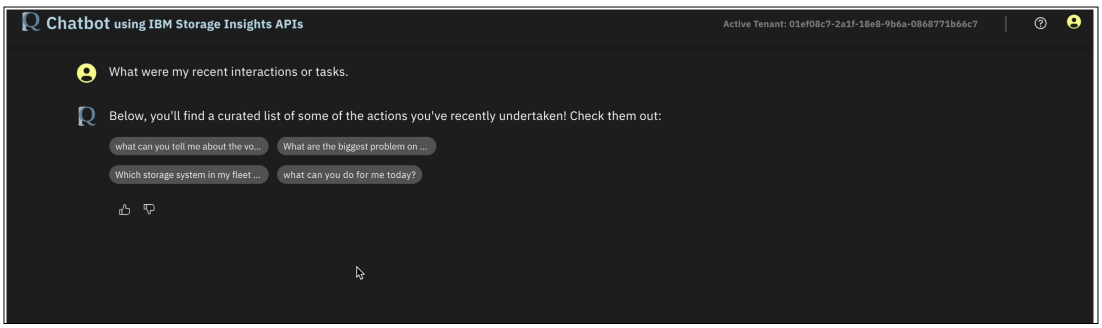


Figure 7-5 Show my recent activity

- Chat Interactions: Users submit queries, receive API-based responses, and refine their requests, as shown in Figure 7-6.

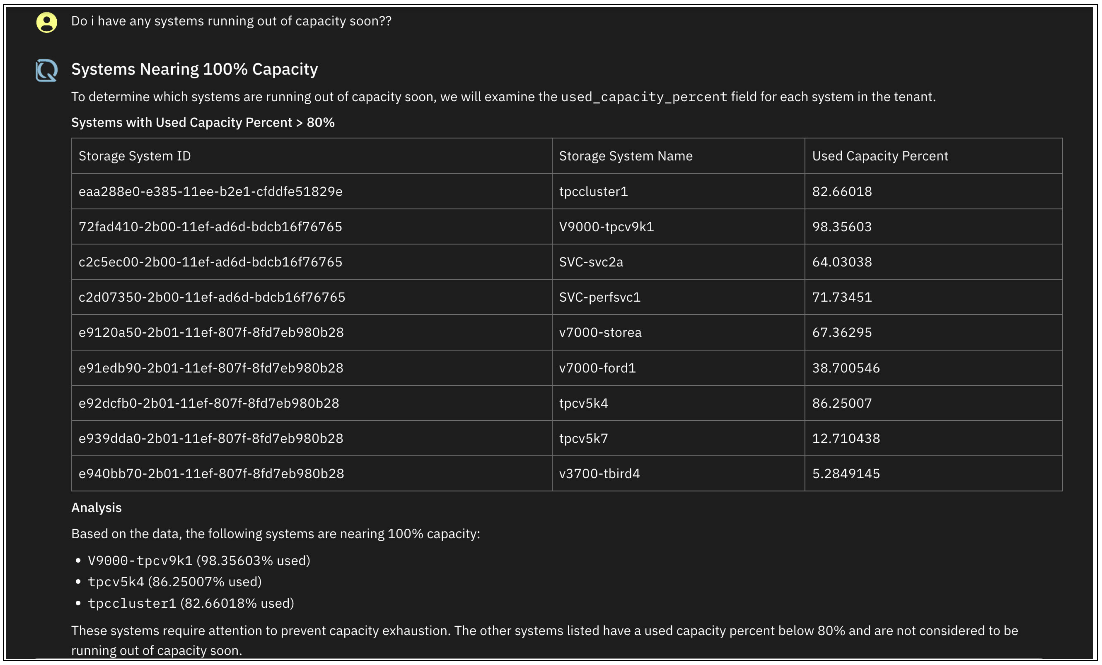


Figure 7-6 Chat interaction example

- Logout: Users can securely log out, which terminates their session, as shown in Figure 7-7.

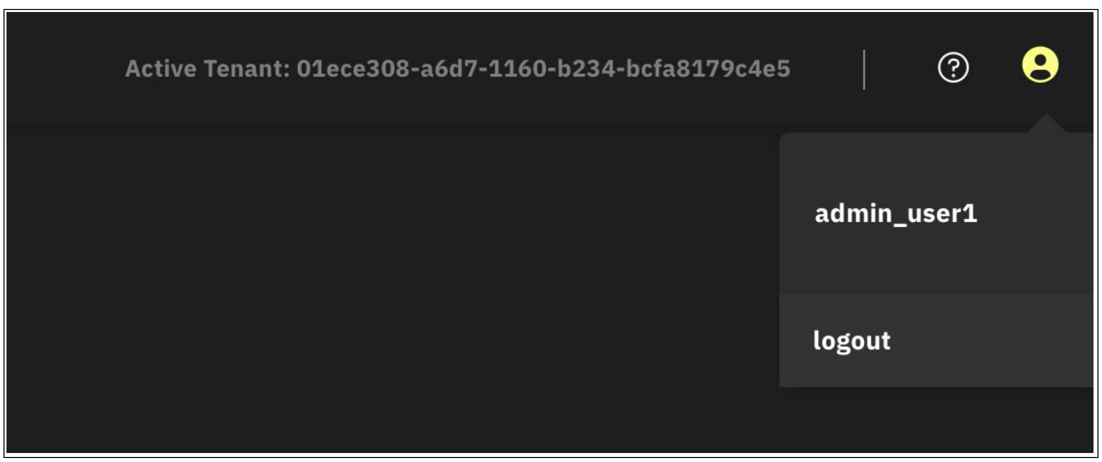


Figure 7-7 Session termination option

7.1.4 Interaction flow

The following steps show an example of using the chatbot:

1. A user initiates a chat by clicking **New Chat**.
2. The routine cards appear. The user selects **Morning Cup of Coffee** or starts a query.
3. The chatbot processes the query and converts it into an API request.
4. A response returns in a structured format.

7.2 Architecture and components

The Observability Chatbot uses a modular, cloud-native architecture that integrates AI, RESTful APIs, and encryption that is based on containerized deployment.

7.2.1 System architecture overview

The architecture follows a layered approach, enabling clear separation of responsibilities across the UI, back-end services, and AI model interaction. At a high level, the chatbot processes user input through a React.js front end, which communicates with a Python-based FastAPI back end. The back end orchestrates natural language understanding, intent and entity extraction by using IBM Granite LLMs, and issues secured REST API calls to IBM Storage Insights. The resulting data is processed, formatted, and returned in a structured conversational response.

User interface layer

The chatbot's front end is a React.js-based web application that is designed to facilitate natural language interactions for seamless observability. It presents retrieved insights in structured formats such as tables, charts, and text summaries, ensuring clear and actionable data visualization. Also, the chatbot supports multi-turn conversations, enabling context-aware dialogues, where users can ask follow-up questions without needing to repeat previous details. This feature enhances usability and efficiency, making storage monitoring and troubleshooting intuitive.

Natural language processing layer

The chatbot leverages IBM watsonx Granite 34B, an advanced large language model (LLM) to accurately interpret user queries and determine the most appropriate IBM Storage Insights API calls. It incorporates three key natural language processing (NLP) components: intent recognition, entity extraction, and context management. Intent recognition enables the chatbot to classify user queries (for example, “Fetch storage alerts”) and map them to the correct API endpoints. Entity extraction ensures the identification of relevant storage-related metadata, such as storage system names, alert severity, and performance metrics, enabling precise API interactions. Also, context management enables the chatbot to retain conversational memory, ensuring that follow-up queries are interpreted correctly within multi-step interactions.

From a technical standpoint, the chatbot implements Greedy Decoding to improve intent classification accuracy and leverages custom prompt engineering to enhance entity recognition. By integrating IBM watsonx Granite 34B and Llama 3 405B parameter models, the chatbot retrieves raw API responses and refines and restructures them into a well-organized, human-readable format. By providing clear, concise, and contextually relevant responses, this approach improves observability workflows and makes storage monitoring more intuitive for users.

Back-end processing layer

The back-end processing layer of the chatbot is built by using FastAPI and is responsible for orchestrating the chatbot's core functions. It ensures seamless interaction between the user interface, NLP engine, and IBM Storage Insights APIs. The back end incorporates session management, enabling it to retain and track ongoing user conversations and context, ensuring continuity in multi-turn interactions. Also, it handles API request processing, where it parses the NLP engine's output, identifies the correct IBM Storage Insights API endpoints, and constructs structured API calls. When data is retrieved from the APIs, the back end performs response formatting, processing raw JSON responses into a structured format, enhancing readability and usability.

IBM Storage Insights API layer

The chatbot directly integrates with IBM Storage Insights APIs, enabling users to fetch real-time observability data that is related to storage performance, alerts, and system status. The chatbot supports multiple API endpoints, including the Storage System Alerts API, which retrieves real-time storage alerts based on severity levels, and the Storage System Metrics API, which provides detailed performance data, usage statistics, and system health metrics. Also, the Tenant Alerts API lists alerts across all storage systems within a tenant, and the Storage System List API fetches a comprehensive list of monitored storage systems.

The API request flow follows a structured process:

1. The user submits a natural language query through the chatbot interface.
2. The NLP engine extracts the user's intent and relevant entities.
3. The back end processes the intent and formulates an API request.
4. The IBM Storage Insights API responds with real-time insights.
5. The back end processes, formats, and displays the response in a structured and readable format.

This seamless integration ensures that users can retrieve actionable insights without needing to manually run API requests, which simplifies observability and monitoring workflows. For the list of all supported IBM Storage Insights APIs, see Table 7-1 on page 65.

Security and authentication layer

The chatbot implements IBM Storage Insights API Key authentication and secure encryption mechanisms to ensure secure access control and data protection during all interactions. To ensure secure access to IBM Storage Insights, the chatbot implements multiple security layers that protect user credentials and data exchanges, aligning with enterprise security standards. Authentication is enforced through OAuth-based authentication, which provides secure access control and prevents unauthorized API calls. Users are required to provide a Tenant ID and API Key during login. On successful authentication, an access token is generated and stored locally, enabling users to interact with the chatbot without repeated authentication. The token remains valid until the user logs out or the back-end server restarts.

Also, all data exchanges between the chatbot and IBM Storage Insights APIs are protected by using TLS/SSL encryption, ensuring end-to-end security and data integrity during communication. For session management, the chatbot automatically handles token expiry. If an expired token is detected, API requests return a 401 error, triggering an automatic logout (see Figure 7-8 on page 65). Then, the chatbot prompts the user to reauthenticate to regain access, preventing unauthorized or stale API requests. These comprehensive security mechanisms ensure a robust, enterprise-grade security framework, safeguarding user data, API credentials, and chatbot interactions while maintaining compliance with IBM's security policies.

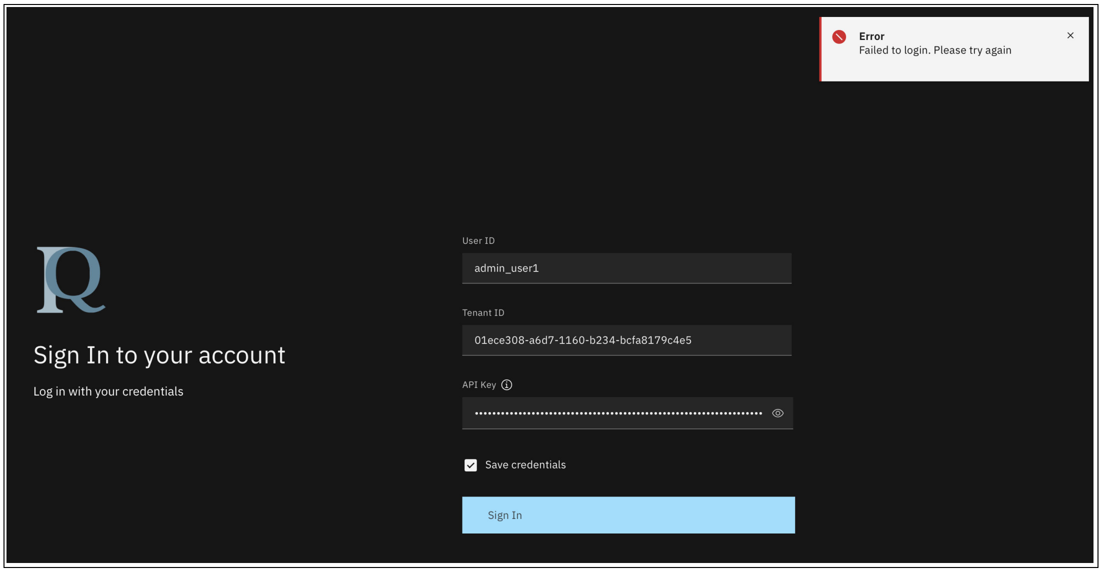


Figure 7-8 Invalid API Key Error

7.3 Supported APIs and capabilities

The chatbot supports the IBM Storage Insights API endpoints that are listed in Table 7-1.

Table 7-1 Supported IBM Storage Insights API endpoints

API endpoint	Function
Storage System Alerts	Fetches alerts based on severity, status, and system type.
Storage System Metrics	Retrieves performance data, capacity usage, and I/O statistics.
Storage System Notifications	Lists system warnings, maintenance events, and status changes.
Storage System List	Provides a list of all monitored storage systems.
Tenant Alerts	Aggregates alerts across all storage systems under a tenant.
Tenant Notifications	Retrieves notifications affecting the entire tenant.
Storage System Volumes	Retrieves volume-specific details like size, provisioning status, and usage.
Storage System Details	Fetches metadata and configuration information for a specific storage system.

Note: This chatbot is a do-it-yourself, generative AI-powered solution with limited API integration. Users can enhance its functionality by integrating additional IBM Storage Insights API endpoints.

7.4 Installation and deployment

This section outlines the prerequisites and step-by-step instructions to install and deploy the Observability Chatbot in a containerized environment.

7.4.1 Prerequisites

Before deploying the chatbot, ensure that the following prerequisites are met:

- IBM watsonx API Key: Required for LLM-powered intent detection and entity recognition.
- IBM Storage Insights API Key: Needed to authenticate external REST API calls.
- Podman: Required for containerized deployment.
- Git: Necessary for cloning the Observability Chatbot repository.
- OpenSSL: Ensures secure HTTPS communication.

To verify the presence of required dependencies, run the following commands:

```
podman info  # Verify Podman installation
openssl version  # Check OpenSSL version
```

7.4.2 Installation steps

To install the Observability Chatbot, complete the following steps:

1. Clone the repository by running the following command:

   ```
   git clone https://github.com/ibmstorage/ibm-storageinsights-chatbot.git
   cd ibm-storageinsights-chatbot/
   ```

2. Generate encryption keys by running the following command:

   ```
   chmod +x generate_key.sh
   ./generate_key.sh
   ```

3. Generate the self-signed certificates by running the following command:

   ```
   chmod +x generate_certificates.sh
   ./generate_certificates.sh
   ```

4. Configure the environment variables:

   a. Open `dockerfile.backend` and update it with the following lines:

   ```
   WATSONX_APIKEY=<your_watsonx_api_key>
   PROJECT_ID=<your_project_id>
   WATSONX_HOSTED_SERVICE=<your_region_endpoint>
   ```

   b. Open `dockerfile.frontend` and update it with the following lines:

   ```
   REACT_APP_BACKEND_BASE_URL=https://<your_host>:9508
   REACT_APP_SECRET_KEY=<your_generated_secret_key>
   ```

5. Deploy the chatbot by running the following command:

   ```
   chmod +x install_si_chatbot.sh
   ./install_si_chatbot.sh
   ```

6. Access the chatbot UI by running the following command:

   ```
   https://<host>:9502/chatbot/
   ```

7.5 Feature engineering

The Observability Chatbot incorporates advanced AI engineering techniques to optimize performance and usability.

7.5.1 Prompt engineering for LLM optimization

The chatbot employs structured prompt engineering techniques to refine LLM-powered responses, which ensures high accuracy in intent detection, entity recognition, and response formatting:

- Contextual prompts:

  The chatbot uses a generalized intent detection prompt to classify user queries into the correct intent category and a generalized entity extraction prompt to identify relevant storage-related entities from user input. These broad prompts enable the chatbot to dynamically extract intent and entities for any user query, making the system highly adaptable.

  However, when an API response is received, context-specific handling is required to ensure that the response is structured, relevant, and human-readable. To achieve this task, the chatbot employs custom contextual prompts that are tailored for different API responses. Each contextual prompt is designed to do the following tasks:

  – Interpret and reformat raw API data for enhanced readability.

  – Summarize key insights based on the API's response structure.

  – Provide structured output formats (tables, bullet points, or textual summaries) for improved user experience.

  For example, when retrieving storage alerts, a specialized prompt restructures the API response to highlight critical alerts first, filter non-essential information, and present actionable insights in a digestible format. These intent-specific contextual prompts ensure that users receive meaningful, well-organized responses instead of raw, unstructured API data.

- Structured response:

  Converts raw API responses into readable, structured insights.

  Example 7-1 shows an example transformation.

Example 7-1 An example transformation

```
Raw API Response:
{
    "alert_id": "1234",
    "severity": "CRITICAL",
    "timestamp": "2024-03-15T10:00:00Z"
}
Processed Response:
- **Alert ID:** 1234
- **Severity:** CRITICAL
- **Timestamp:** 15 March 2024, 10:00 AM
```

Example 7-2 on page 68 depicts an example prompt structure for handling responses of storage alerts.

Example 7-2 An example prompt structure for handling responses of storage alerts

```
Response Generation Instruction for Storage System Alert

"""Below document provides details about alerts that were
received for a storage system. The answer to the user query is present in the
document. Answer questions only
about alerts for a storage system. Your answer should always include the following
fields: alert ID, parent resource
name, alert name, severity, occurrence time, source of the alert, creator of the
alert in addition to whatever fields
you choose to add. If you are asked about notifications, details about a system,
the below document does not provide that
answer."""
```

Figure 7-9 shows an example of a structured list response.

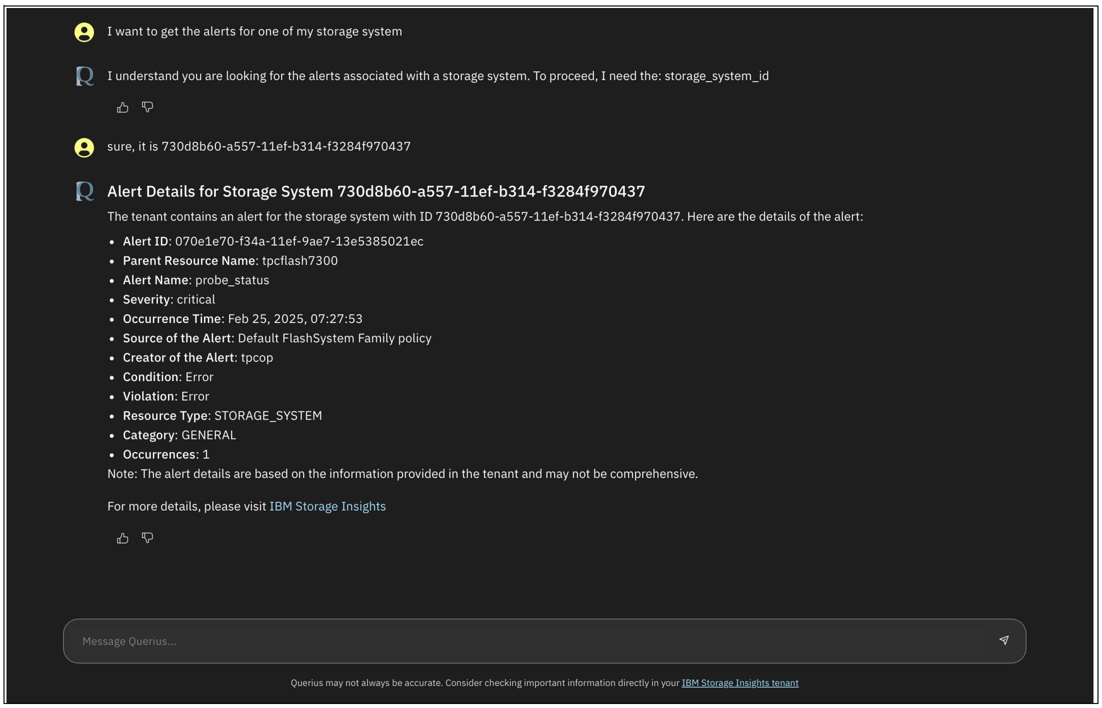


Figure 7-9 Structured list response

- Error-Handling Prompts

  Detects unrecognized intents and guides users toward better query formulation.

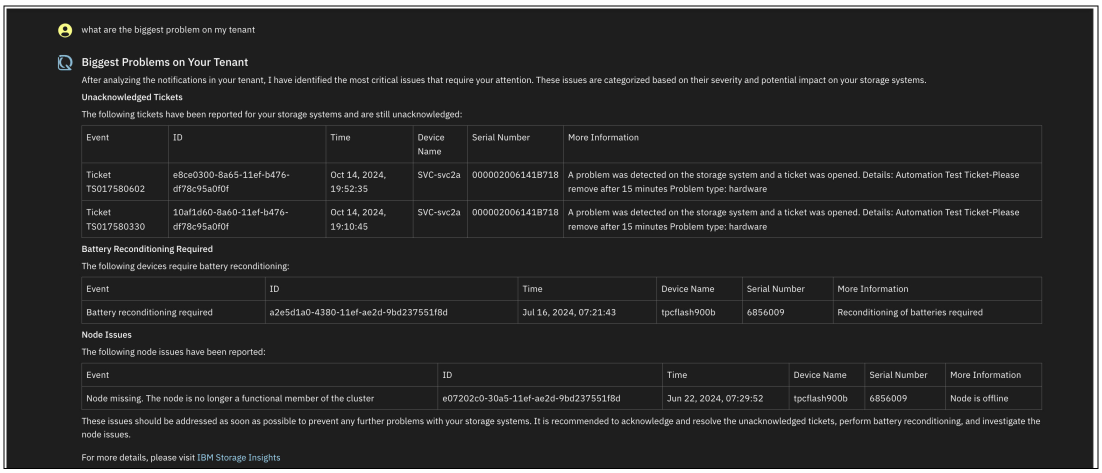
7.5.2 Response handling

The chatbot is designed to generate thoughtful and structured responses by following a response generation template that ensures clarity, correctness, and contextual relevance. The response handling process begins with the LLM evaluating the user's query, extracting intent and relevant entities, and then determining how the raw API response should be structured before presenting it to the user.

The response template enforces several key principles:

- Context awareness: Ensures that the response is grounded in the specific tenant's data. If the API returns an empty tenant response, the chatbot clearly informs the user that no relevant data is available
- Explanatory responses: If an answer requires an explanation, the chatbot first provides context or reasoning before delivering the final answer.
- Markdown formatting: Responses are structured by using markdown elements (for example, headings, bold text, bullet points, and code blocks) for enhanced readability.
- Dynamic formatting: The chatbot chooses between text, tables, or lists depending on the nature of the API response and the user query.
- Mathematical constraints: The chatbot explicitly avoids performing math calculations and instead informs users that mathematical operations are not supported.

Figure 7-10 shows a structured response.


what are the biggest problem on my tenant

Biggest Problems on Your Tenant

After analyzing the notifications in your tenant, I have identified the most critical issues that require your attention. These issues are categorized based on their severity and potential impact on your storage systems.

Unacknowledged Tickets

The following tickets have been reported for your storage systems and are still unacknowledged:

Event	ID	Time	Device Name	Serial Number	More Information
Ticket TS017580602	e8ce0300-8a65-11ef-b476-df78c95a0f0f	Oct 14, 2024, 19:52:35	SVC-svc2a	000002006141B718	A problem was detected on the storage system and a ticket was opened. Details: Automation Test Ticket-Please remove after 15 minutes Problem type: hardware
Ticket TS017580330	10af1d60-8a60-11ef-b476-df78c95a0f0f	Oct 14, 2024, 19:10:45	SVC-svc2a	000002006141B718	A problem was detected on the storage system and a ticket was opened. Details: Automation Test Ticket-Please remove after 15 minutes Problem type: hardware

Battery Reconditioning Required

The following devices require battery reconditioning:

Event	ID	Time	Device Name	Serial Number	More Information
Battery reconditioning required	a2e5d1a0-4380-11ef-ae2d-9bd237551f8d	Jul 16, 2024, 07:21:43	tpcflash900b	6856009	Reconditioning of batteries required

Node Issues

The following node issues have been reported:

Event	ID	Time	Device Name	Serial Number	More Information
Node missing. The node is no longer a functional member of the cluster	e07202c0-30a5-11ef-ae2d-9bd237551f8d	Jun 22, 2024, 07:29:52	tpcflash900b	6856009	Node is offline

These issues should be addressed as soon as possible to prevent any further problems with your storage systems. It is recommended to acknowledge and resolve the unacknowledged tickets, perform battery reconditioning, and investigate the node issues.

For more details, please visit IBM Storage Insights

Figure 7-10 Structured response

- Visualization support: Displays alerts in tabular formats and supports exporting results in csv format, as shown in Figure 7-11.

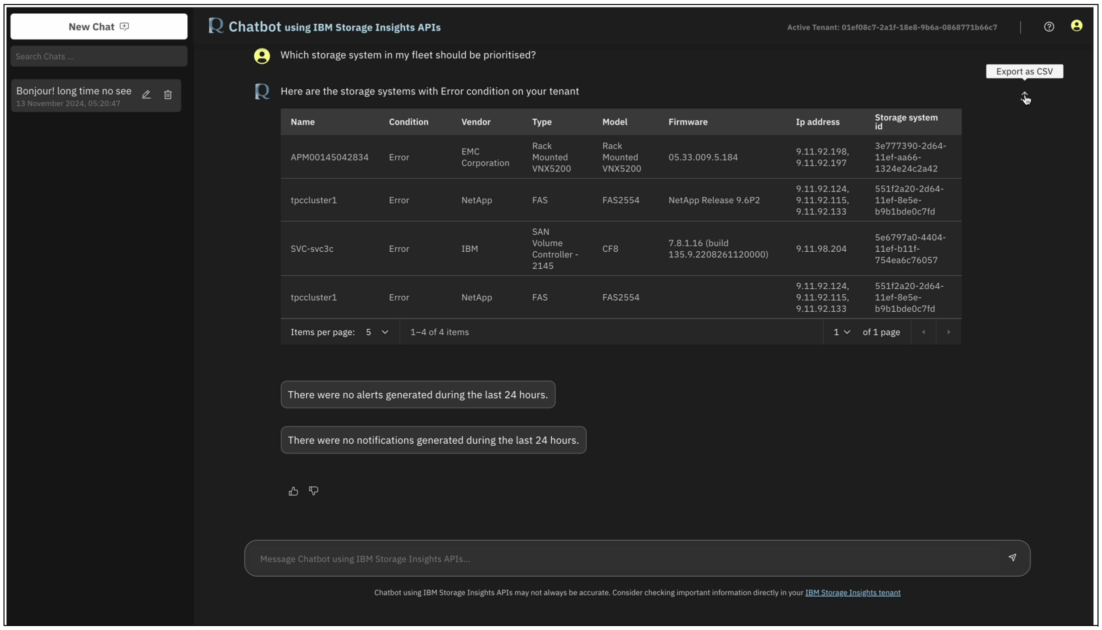


Figure 7-11 Tabular response

7.5.3 Error handling and incomplete intent management

Here are the functions for error handling and incomplete intent:

- Graceful failures:
  - Implements fallback responses when an LLM cannot determine intent.
  - Prevents repetitive failed responses by setting an unresolved intent threshold.

  Figure 7-12 shows a graceful exit scenario.

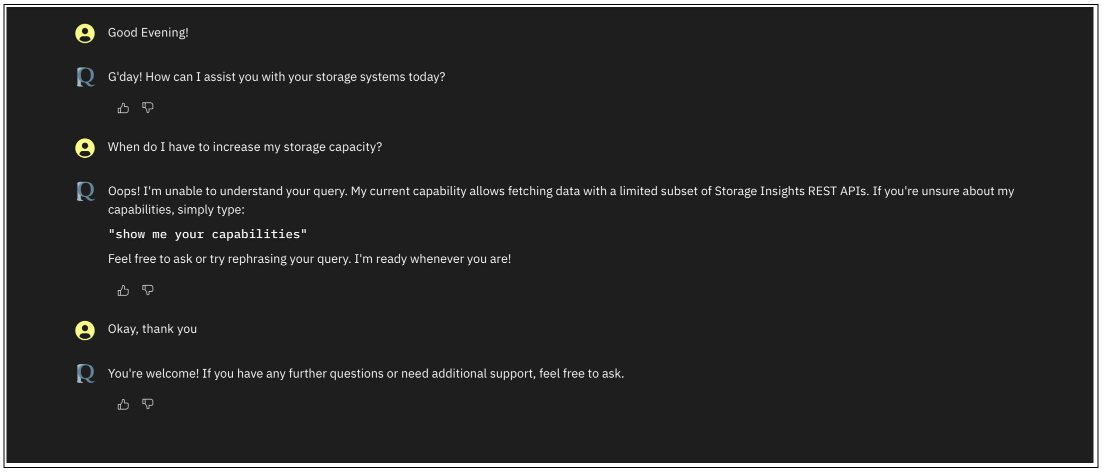


Figure 7-12 Graceful exit scenario

- Continuous intent tracking: If an incomplete intent is detected, the chatbot tracks missing details and asks clarifying questions.

  Figure 7-13 shows an example of continuous intent tracking.

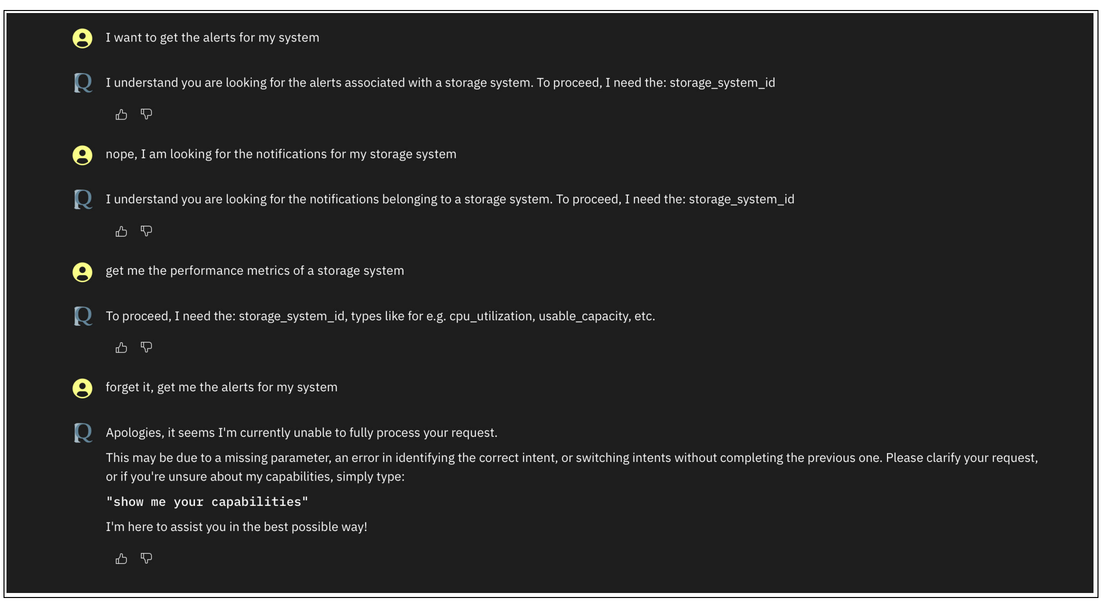


Figure 7-13 Continuous intent tracking and exit on unfulfillment of entity

7.5.4 Performance optimizations

Here are performance optimizations for the chatbot:

- Pre-caching frequently requested API calls

  To enhance performance and reduce redundant API requests, the chatbot implements a pre-caching mechanism for frequently accessed IBM Storage Insights API calls. This optimization stores previously fetched responses in a temporary cache, enabling the chatbot to serve repeated queries instantly without making more API calls. By leveraging caching, response times are improved, especially for high-frequency queries such as storage system lists, recent alerts, and performance metrics. This approach enhances the user experience by reducing latency, and minimizes unnecessary load on the IBM Storage Insights API.

- Conversation history and threading

  To enhance chat continuity and user experience, the chatbot retains historical chat interactions, enabling users to resume previous conversations across sessions without losing context. Each conversation is stored as a new thread in the database, enabling users to revisit and continue past interactions from the Chat History window. This function ensures that users do not have to re-enter previous queries, improving efficiency and reducing redundancy in API calls.

  The chatbot uses SQLite for persistent chat storage, ensuring that interactions remain accessible even after a session ends. This structure optimizes memory usage and response retrieval, reducing latency in multi-turn conversations. To maintain data integrity and optimize storage, only valid chatbot responses (with a 200 OK status) are stored. This approach prevents unnecessary retention of incomplete or erroneous responses, ensuring that the conversation history remains relevant, accurate, and efficient.

7.6 Challenges and solutions

The development of the Observability Chatbot required addressing several challenges related to accuracy, conversational flow, and API performance optimization. These challenges were systematically mitigated by leveraging IBM watsonx capabilities and implementing custom engineering enhancements to improve the chatbot’s reliability and usability.

The following are key challenges and solutions:

- LLM Hallucinations: One of the key challenges was LLM hallucinations, where the chatbot occasionally generated responses that included incorrect or non-existent entities. To enhance accuracy, entity validation layers were incorporated into the chatbot architecture. These validation mechanisms cross-check extracted entities against IBM Storage Insights metadata, ensuring that only verified data is used in API requests. This approach reduced false positives, improving the chatbot’s ability to deliver precise and contextually relevant insights.
- Limited Intent Chaining: Another challenge was limited intent chaining, which impacted the chatbot’s ability to handle multi-step interactions effectively. Traditional LLMs require explicit re-specification of entities in each query, limiting the natural flow of conversations. A custom state-tracking mechanism was implemented to address this challenge, enabling the chatbot to retain conversational memory and dynamically link follow-up queries to previous responses. This enhancement improved the fluidity of interactions, allowing users to refine or expand their queries without repetition.
- API Rate Limits: The IBM Storage Insights API rate limits presented a constraint because excessive API requests could lead to performance degradation or temporary service

unavailability. Caching mechanisms were introduced to store frequently accessed API responses, optimizing API utilization, reducing redundant calls, and ensuring efficient data retrieval. This solution improved response times and enhanced the chatbot's ability to provide high-priority observability insights without exceeding API quotas.

7.6.1 Advancements with Granite 3 models

With the launch of the Granite 3 models, IBM's latest advancements in LLM technology now include function calling, improved intent chaining, and enhanced summarization capabilities for large API responses. These innovations enable seamless multi-step interactions, reducing the need for extensive state-tracking logic. Furthermore, the new models can summarize extensive API outputs into concise, structured insights, enhancing the overall user experience.

7.6.2 Future roadmap

As part of the long-term roadmap, future iterations of the chatbot will migrate to IBM native watsonx Granite models, eliminating the dependency on external LLMs such as Llama. This approach will further strengthen data security, integration efficiency, and AI-driven observability capabilities, aligning with IBM's strategic direction for enterprise AI solutions.

7.7 Conclusion

The Observability Chatbot for IBM Storage Insights leverages LLMs, REST API integration, and advanced prompt engineering to enhance real-time storage observability. With stateful conversations, structured API responses, and security measures, it serves as a powerful AI-driven assistant for IT administrators and SREs.

This chatbot represents the next generation of AI-powered observability solutions, offering a simplified and intelligent way to manage enterprise storage infrastructure.

> **Note:** For more information the Observability Chatbot for IBM Storage Insights, you can watch the following video.

8

Performance anomaly detection

One of the important uses of Artificial Intelligence within IBM Storage Insights is that it helps resolve performance issues faster by allowing the support team to leverage the power of AI to detect performance anomalies. This is possible as IBM Storage Insights stores the performance data eliminating the need for clients to upload logs. This also reduces the time to resolution for performance issues which could otherwise take more time.

This chapter briefly discusses the performance anomaly detection in IBM Storage Insights and has the following sections:

Note: This feature is available for IBM Support teams to resolve the tickets faster.

8.1 Introduction to performance anomaly detection

Imagine a storage administrator overseeing a mission-critical environment where downtime spells significant financial repercussions. Suddenly, a crippling performance issue emerges, inundating them with a barrage of alerts and frustrated user complaints. Traditionally, resolving such a crisis would demand painstakingly sifting through disparate logs, meticulously tracing the root cause, and engaging in lengthy back-and-forth with the support team - a process that could easily consume hours, if not days.

IBM Storage Insights can be the savior whenever a performance issue is observed. As soon as a performance issue is reported, the storage administrator can raise a support ticket directly from the Storage Insights interface, eliminating the need to manually upload logs. IBM Storage Insights, on seeing an open ticket, starts the AI based performance anomaly detection in the background and keeps the result ready for the IBM Support personnel. The performance anomaly is identified by looking through the performance pattern observed across more than 200 performance metrics. The IBM Support personnel now have access to a heatmap showcasing if an anomaly is observed across the categorized performance metrics. This significantly reduces the time to pinpoint a performance issue. A snippet of such a view is shown in Figure 8-1.

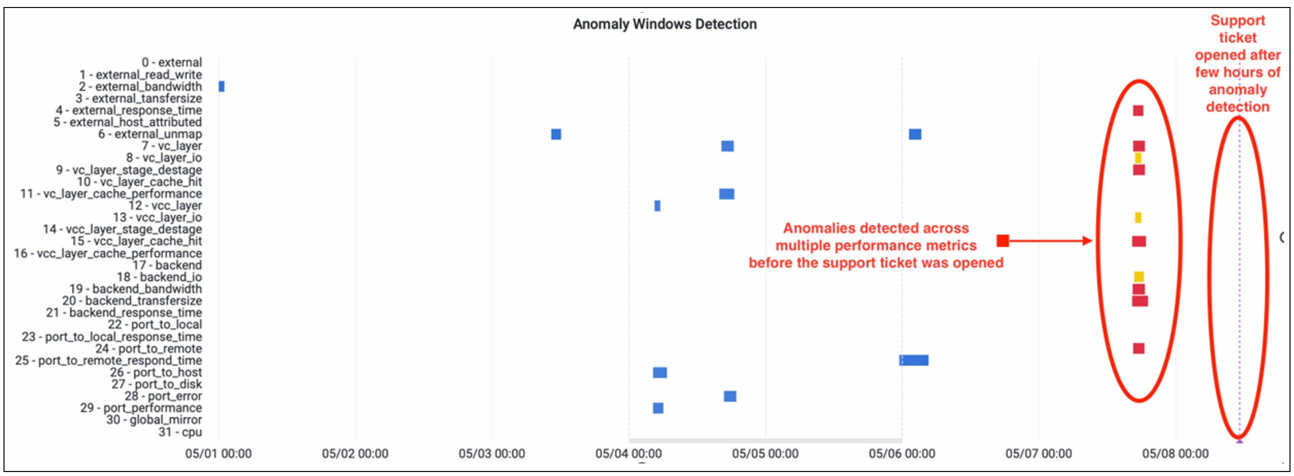


Figure 8-1 Performance anomaly detection

> **Note:** Performance anomaly detection capability is available to the IBM Support users only. The image shown above is an illustration of the anomaly detection capability

As illustrated in Figure 8-1, the system immediately grants the IBM Support team access to your storage logs, performance data, and historical insights. This significantly accelerates issue resolution, enabling faster solutions compared to systems lacking IBM Storage Insights. AI for storage plays a crucial role in support for troubleshooting, monitoring, and optimizing storage systems.

Leveraging AI-powered anomaly detection, IBM Storage Insights analyzes historical performance patterns to rapidly pinpoint the root cause. Armed with this data, the IBM Support team can identify the issue in a fraction of the manual effort. They utilize advanced diagnostic tools like Performance Anomaly Detection to deliver immediate recommendations or actions, dramatically accelerating the resolution process.

8.2 How IBM Support team leverages artificial intelligence in IBM Storage Insights

Here is the summary of how IBM Support team leverages artificial intelligence in IBM Storage Insights:

1. A support ticket is opened using IBM Storage Insights or by other means of reporting an issue.
2. If the system where the issue is reported is configured to use IBM Storage Insights, an automated AI based anomaly detection is triggered for the system.
3. The AI capability built by IBM research and currently used by IBM Storage support through IBM Storage Insights provides a summary of the anomalies observed.
4. The IBM Support representative can further use the anomaly detection capability to deep dive into individual performance data and isolate the root cause for the reported issue.
5. Once identified the action plan to address the issue is then shared and executed.

Overall, performance anomaly detection capability helps to reduce the time taken to resolve complex performance issues.

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this paper.

IBM Redbooks

The following IBM Redbooks publications provide additional information about the topic in this document. Note that some publications referenced in this list might be available in softcopy only.

- *Maximize Your Storage Investment: Use IBM Storage Insights for Observability and Artificial Intelligence for IT Operations*, SG24-8584

You can search for, view, download or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

Online resources

These websites are also relevant as further information sources:

- IBM Storage Insights and IBM Spectrum Control

  `https://www.ibm.com/products/storage-insights`

- Videos for IBM Storage Insights

  `https://www.ibm.com/docs/en/storage-insights?topic=media-videos`

- Blogs for IBM Storage Insights

  `https://www.ibm.com/docs/en/storage-insights?topic=media-blogs`

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

REDP-5755-00

ISBN

Printed in U.S.A.